(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 033 401 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
06.09.2000 Bulletin 2000/36

(51) Int. Cl.[7]: **C12N 15/12**, C07K 14/47, C07K 16/18, G06F 17/30

(21) Application number: **00200610.4**

(22) Date of filing: **21.02.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **26.02.1999 US 122487 P**

(71) Applicant: **GENSET**
75008 Paris (FR)

(72) Inventors:
, Dumas Milne Edwards, Jean-Baptiste
75006 Paris (FR)
, Duclert, Aymeric
94100 Saint-Maur (FR)
, Giordano, Jean-Yves
75018 Paris (FR)

(74) Representative:
Brasnett, Adrian Hugh et al
Mewburn Ellis
York House
23 Kingsway
London WC2B 6HP (GB)

Remarks:
THE COMPLETE DOCUMENT INCLUDING REFERENCE TABLES AND THE SEQUENCE LISTING IS AVAILABLE ON CD-ROM FROM THE EUROPEAN PATENT OFFICE, VIENNA SUB-OFFICE.

(54) Expressed sequence tags and encoded human proteins

(57) The sequences of 5' ESTs derived from mRNAs encoding secreted proteins are disclosed. The 5' ESTs may be to obtain cDNAs and genomic DNAs corresponding to the 5' ESTs. The 5' ESTs may also be used in diagnostic, forensic, gene therapy, and chromosome mapping procedures. Upstream regulatory sequences may also be obtained using the 5' ESTs. The 5' ESTs may also be used to design expression vectors and secretion vectors.

EP 1 033 401 A2

Background of the Invention

[0001] The estimated 50,000-100,000 genes scattered along the human chromosomes offer tremendous promise for the understanding, diagnosis, and treatment of human diseases. In addition, probes capable of specifically hybridizing to loci distributed throughout the human genome find applications in the construction of high resolution chromosome maps and in the identification of individuals.

[0002] In the past, the characterization of even a single human gene was a painstaking process, requiring years of effort. Recent developments in the areas of cloning vectors, DNA sequencing, and computer technology have merged to greatly accelerate the rate at which human genes can be isolated, sequenced, mapped, and characterized. Cloning vectors such as yeast artificial chromosomes (YACs) and bacterial artificial chromosomes (BACs) are able to accept DNA inserts ranging from 300 to 1000 kilobases (kb) or 100-400 kb in length respectively, thereby facilitating the manipulation and ordering of DNA sequences distributed over great distances on the human chromosomes. Automated DNA sequencing machines permit the rapid sequencing of human genes. Bioinformatics software enables the comparison of nucleic acid and protein sequences, thereby assisting in the characterization of human gene products.

[0003] Currently, two different approaches are being pursued for identifying and characterizing the genes distributed along the human genome. In one approach, large fragments of genomic DNA are isolated, cloned, and sequenced. Potential open reading frames in these genomic sequences are identified using bioinformatics software. However, this approach entails sequencing large stretches of human DNA which do not encode proteins in order to find the protein encoding sequences scattered throughout the genome. In addition to requiring extensive sequencing, the bioinformatics software may mischaracterize the genomic sequences obtained. Thus, the software may produce false positives in which non-coding DNA is mischaracterized as coding DNA or false negatives in which coding DNA is mislabeled as non-coding DNA.

[0004] An alternative approach takes a more direct route to identifying and characterizing human genes. In this approach, complementary DNAs (cDNAs) are synthesized from isolated messenger RNAs (mRNAs) which encode human proteins. Using this approach, sequencing is only performed on DNA which is derived from protein coding portions of the genome. Often, only short stretches of the cDNAs are sequenced to obtain sequences called expressed sequence tags (ESTs). The ESTs may then be used to isolate or purify extended cDNAs which include sequences adjacent to the EST sequences. The extended cDNAs may contain all of the sequence of the EST which was used to obtain them or only a portion of the sequence of the EST which was used to obtain them. In addition, the extended cDNAs may contain the full coding sequence of the gene from which the EST was derived or, alternatively, the extended cDNAs may include portions of the coding sequence of the gene from which the EST was derived. It will be appreciated that there may be several extended cDNAs which include the EST sequence as a result of alternate splicing or the activity of alternative promoters. Alternatively, ESTs having partially overlapping sequences may be identified and contigs comprising the consensus sequences of the overlapping ESTs may be identified.

[0005] In the past, these short EST sequences were often obtained from oligo-dT primed cDNA libraries. Accordingly, they mainly corresponded to the 3' untranslated region of the mRNA. In part, the prevalence of EST sequences derived from the 3' end of the mRNA is a result of the fact that typical techniques for obtaining cDNAs, are not well suited for isolating cDNA sequences derived from the 5' ends of mRNAs. (Adams et al., *Nature* **377**:3-174, 1996, Hillier et al., *Genome Res.* **6**:807-828, 1996).

[0006] In addition, in those reported instances where longer cDNA sequences have been obtained, the reported sequences typically correspond to coding sequences and do not include the full 5' untranslated region (5'UTR) of the mRNA from which the cDNA is derived. 5'UTRs are often involved in the regulation of gene expression, by affecting either the stability or translation of mRNAs. Indeed, 5'UTRs may contain several features known to affect the initiation of translation: (i) the distance between the cap structure and the initiation codon, (ii) the presence of *cis-acting* elements which may be either linear sequences such as polypyrimidine tracts (Kaspar *et al, J. Biol. Chem.* **267**, 508-514, 1992; Severson *et al., Eur J Biochem* **229**:426-32, 1995) or secondary structures such as IREs (Rouault and Klausner, *Curr Top Cell Regul***35**:1-19, 1997), and (iii) upstream open reading frames or uORFs (Geballe and Morris, *Trends Biochem Sci* **19**:159-64, 1994). Thus, regulation of gene expression may be achieved through the use of alternative 5'UTRs. For instance, the translation of the tissue inhibitor of metalloprotease mRNA is enhanced in mitogenically activated cells through modification of the start codon of an uORF in its 5'UTR using an alternative promoter (Waterhouse *et al, J Biol Chem.* **265**:5585-9. 1990). Furthermore, modification of 5'UTR through mutation, insertion or translocation events may even be implied in pathogenesis. For instance, the fragile X syndrome, the most common cause of inherited mental retardation, is partly due to an insertion of multiple CGG trinucleotides in the 5'UTR of the fragile X mRNA resulting in the inhibition of protein synthesis via ribosome stalling (Feng *et al, Science* **268**:731-4, 1995). An aberrant mutation in regions of the 5'UTR known to inhibit translation of the proto-oncogene *c-myc* was shown to result in upregulation of C-myc protein levels in cells derived from patients with multiple myelomas (Willis *et al, Curr Top Microbiol Immunol* **224**:269-76, 1997). However, the use of oligo-dT primed cDNA libraries does not allow the isolation of complete 5'UTRs since such obtained incomplete sequences may not include the first exon of the mRNA, particularly in situations where the first exon is short. Furthermore, they may not include some exons, often short ones, which are located upstream of splicing sites. Thus, there is a need to obtain sequences derived from the 5' ends of mRNAs.

[0007] While many sequences derived from human chromosomes have practical applications, approaches based

on the identification and characterization of those chromosomal sequences which encode a protein product are particularly relevant to diagnostic and therapeutic uses. In some instances, the sequences used in such therapeutic or diagnostic techniques may be sequences which encode proteins which are secreted from the cell in which they are synthesized, as well as the secreted proteins themselves, are particularly valuable as potential therapeutic agents. Such proteins are often involved in cell to cell communication and may be responsible for producing a clinically relevant response in their target cells. In fact, several secretory proteins, including tissue plasminogen activator, G-CSF, GM-CSF, erythropoietin, human growth hormone, insulin, interferon-α, interferon-β, interferon-γ, and interleukin-2, are currently in clinical use. These proteins are used to treat a wide range of conditions, including acute myocardial infarction, acute ischemic stroke, anemia, diabetes, growth hormone deficiency, hepatitis, kidney carcinoma, chemotherapy-induced neutropenia and multiple sclerosis. For these reasons, extended cDNAs encoding secreted proteins or portions thereof represent a valuable source of therapeutic agents. Thus, there is a need for the identification and characterization of secreted proteins and the nucleic acids encoding them.

[0008] In addition to being therapeutically useful themselves, secretory proteins include short peptides, called signal peptides, at their amino termini which direct their secretion. These signal peptides are encoded by the signal sequences located at the 5' ends of the coding sequences of genes encoding secreted proteins. These signal peptides can be used to direct the extracellular secretion of any protein to which they are operably linked. In addition, portions of the signal peptides called membrane-translocating sequences, may also be used to direct the intracellular import of a peptide or protein of interest. This may prove beneficial in gene therapy strategies in which it is desired to deliver a particular gene product to cells other than the cell in which it is produced. Signal sequences encoding signal peptides also find application in simplifying protein purification techniques. In such applications, the extracellular secretion of the desired protein greatly facilitates purification by reducing the number of undesired proteins from which the desired protein must be selected. Thus, there exists a need to identify and characterize the 5' portions of the genes for secretory proteins which encode signal peptides.

[0009] Sequences coding for non-secreted proteins may also find application as therapeutics or diagnostics. In particular, such sequences may be used to determine whether an individual is likely to express a detectable phenotype, such as a disease, as a consequence of a mutation in the coding sequence for a non-secreted protein or for a secreted protein. In instances where the individual is at risk of suffering from a disease or other undesirable phenotype as a result of a mutation in such a coding sequence, the undesirable phenotype may be corrected by introducing a normal coding sequence using gene therapy. Alternatively, if the undesirable phenotype results from overexpression of the protein encoded by the coding sequence, expression of the protein may be reduced using antisense or triple helix based strategies.

[0010] The secreted or non-secreted human polypeptides encoded by the coding sequences may also be used as therapeutics by administering them directly to an individual having a condition, such as a disease, resulting from a mutation in the sequence encoding the polypeptide. In such an instance, the condition can be cured or ameliorated by administering the polypeptide to the individual.

[0011] In addition, the secreted or non-secreted human polypeptides or portions thereof may be used to generate antibodies useful in determining the tissue type or species of origin of a biological sample. The antibodies may also be used to determine the cellular localization of the secreted or non-secreted human polypeptides or the cellular localization of polypeptides which have been fused to the human polypeptides. In addition, the antibodies may also be used in immunoaffinity chromatography techniques to isolate, purify, or enrich the human polypeptide or a target polypeptide which has been fused to the human polypeptide.

[0012] Public information on the number of human genes for which the promoters and upstream regulatory regions have been identified and characterized is quite limited. In part, this may be due to the difficulty of isolating such regulatory sequences. Upstream regulatory sequences such as transcription factor binding sites are typically too short to be utilized as probes for isolating promoters from human genomic libraries. Recently, some approaches have been developed to isolate human promoters. One of them consists of making a CpG island library (Cross *et al.*, , *Nature Genetics* 6: 236-244, 1994). The second consists of isolating human genomic DNA sequences containing SpeI binding sites by the use of SpeI binding protein. (Mortlock *et al., Genome Res.* 6:327-335, 1996). Both of these approaches have their limits due to a lack of specificity or because they are not universally applicable since only a limited number of promoters have either a CpG island or a Spe I recognition site and because Spe I binding sites are not specifically found in promoter regions. Thus, there exists a need to identify and systematically characterize the 5' portions of the genes.

[0013] The present 5' ESTs may be used to efficiently identify and isolate 5'UTRs and upstream regulatory regions which control the location, developmental stage, rate, and quantity of protein synthesis, as well as the stability of the mRNA. Once identified and characterized, these regulatory regions may be utilized in gene therapy or protein purification schemes to obtain the desired amount and locations of protein synthesis or to inhibit, reduce, or prevent the synthesis of undesirable gene products.

[0014] In addition, ESTs containing the 5' ends of protein genes may include sequences useful as probes for chromosome mapping and the identification of individuals. Thus, there is a need to identify and characterize the sequences upstream of the 5' coding sequences of genes.

Summary of the Invention

[0015] The present invention relates to purified, isolated, or enriched 5' ESTs which include sequences derived from the authentic 5' ends of their corresponding mRNAs. The term "corresponding mRNA" refers to the mRNA which

was the template for the cDNA synthesis which produced the 5' EST. These sequences will be referred to hereinafter as "5' ESTs." The present invention also includes purified, isolated or enriched nucleic acids comprising contigs assembled by determining a consensus sequences from a plurality of ESTs containing overlapping sequences. These contigs will be referred to herein as "consensus contigated ESTs."

[0016] As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual 5' EST clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an approximately 10^4à-10^6à fold purification of the native message. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated.

[0017] As used herein, the term "isolated" requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide present in a living animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated.

[0018] As used herein, the term "enriched" means that the 5' EST is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the 5' ESTs will represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the present invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. Preferably, the enriched 5' ESTs represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More preferably, the enriched 5' ESTs represent 50% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. In a highly preferred embodiment, the enriched 5' ESTs represent 90% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules.

[0019] "Stringent", "moderate," and "low" hybridization conditions are as defined below.

[0020] The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude post-expression modifications of polypeptides, for example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

[0021] As used interchangeably herein, the terms "nucleic acids", "oligonucleotides", and "polynucleotides" include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term "nucleotide" as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT publication No. WO 95/04064. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art.

[0022] The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another be virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, L., *Biochemistry*, 4thà edition, 1995).

[0023] The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide", "complementary nucleic acid" and "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind. Preferably, a "complementary" sequence is a sequence which an A at each position where there is a T on the opposite strand, a T at each position where there is an A on the opposite strand, a G at each position where there is a C on

the opposite strand and a C at each position where there is a G on the opposite strand.

[0024] Thus, 5' ESTs in cDNA libraries in which one or more 5' ESTs make up 5% or more of the number of nucleic acid inserts in the backbone molecules are "enriched recombinant 5' ESTs" as defined herein. Likewise, 5' ESTs in a population of plasmids in which one or more 5' ESTs of the present invention have been inserted such that they represent 5% or more of the number of inserts in the plasmid backbone are "enriched recombinant 5' ESTs" as defined herein. However, 5' ESTs in cDNA libraries in which 5' ESTs constitute less than 5% of the number of nucleic acid inserts in the population of backbone molecules, such as libraries in which backbone molecules having a 5' EST insert are extremely rare, are not "enriched recombinant 5' ESTs."

[0025] In some embodiments, the present invention relates to 5' ESTs which are derived from genes encoding secreted proteins. As used herein, a "secreted" protein is one which, when expressed in a suitable host cell, is transported across or through a membrane, including transport as a result of signal peptides in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g. soluble proteins), or partially (e.g. receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum.

[0026] Such 5' ESTs include nucleic acid sequences, called signal sequences, which encode signal peptides which direct the extracellular secretion of the proteins encoded by the genes from which the 5' ESTs are derived. Generally, the signal peptides are located at the amino termini of secreted proteins.

[0027] Secreted proteins are translated by ribosomes associated with the "rough" endoplasmic reticulum. Generally, secreted proteins are co-translationally transferred to the membrane of the endoplasmic reticulum. Association of the ribosome with the endoplasmic reticulum during translation of secreted proteins is mediated by the signal peptide. The signal peptide is typically cleaved following its co-translational entry into the endoplasmic reticulum. After delivery to the endoplasmic reticulum, secreted proteins may proceed through the Golgi apparatus. In the Golgi apparatus, the proteins may undergo post-translational modification before entering secretory vesicles which transport them across the cell membrane.

[0028] The 5' ESTs of the present invention have several important applications. For example, they may be used to obtain and express cDNA clones which include the full protein coding sequences of the corresponding gene products, including the authentic translation start sites derived from the 5' ends of the coding sequences of the mRNAs from which the 5' ESTs are derived. These cDNAs will be referred to hereinafter as "full-length cDNAs." These cDNAs may also include DNA derived from mRNA sequences upstream of the translation start site. The full-length cDNA sequences may be used to express the proteins corresponding to the 5' ESTs. As discussed above, secreted proteins and non-secreted proteins may be therapeutically important. Thus, the proteins expressed from the cDNAs may be useful in treating or controlling a variety of human conditions. The 5' ESTs may also be used to obtain the corresponding genomic DNA. The term "corresponding genomic DNA" refers to the genomic DNA which encodes the mRNA from which the 5' EST was derived.

[0029] Alternatively, the 5' ESTs may be used to obtain and express extended cDNAs encoding portions of the protein. In the case of secreted proteins, the portions may comprise the signal peptides of the secreted proteins or the mature proteins generated when the signal peptide is cleaved off.

[0030] The present invention includes isolated, purified, or enriched "EST-related nucleic acids." The terms "isolated", "purified" or "enriched" have the meanings provided above. As used herein, the term "EST-related nucleic acids" means the nucleic acids of SEQ ID NOs: 24-4100 and 8178-36681, extended cDNAs obtainable using the nucleic acids of SEQ ID NOs: 24-4100 and 8178-36681, full-length cDNAs obtainable using the nucleic acids of SEQ ID NOs: 24-4100 and 8178-36681 or genomic DNAs obtainable using the nucleic acids of SEQ ID NOs: 24-4100 and 8178-36681. The present invention also includes the sequences complementary to the EST-related nucleic acids.

[0031] The present invention also includes isolated, purified, or enriched "fragments of EST-related nucleic acids." The terms "isolated", "purified" and "enriched" have the meanings described above. As used herein the term "fragments of EST-related nucleic acids" means fragments comprising at least 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 consecutive nucleotides of the EST-related nucleic acids to the extent that fragments of these lengths are consistent with the lengths of the particular EST-related nucleic acids being referred to. The present invention also includes the sequences complementary to the fragments of the EST-related nucleic acids.

[0032] The present invention also includes isolated, purified, or enriched "positional segments of EST-related nucleic acids." The terms "isolated", "purified", or "enriched" have the meanings provided above. As used herein, the term "positional segments of EST-related nucleic acids" includes segments comprising nucleotides 1-25, 26-50, 51-75, 76-100, 101-125, 126-150, 151-175, 176-200, 201-225, 226-250, 251-300, 301-325, 326-350, 351-375, 376-400, 401-425, 426-450, 451-475, 476-500, 501-525, 526-550, 551-575, 576-600 and 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referred to. The term "positional segments of EST-related nucleic acids also includes segments comprising nucleotides 1-50, 51-100, 101-150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450, 450-500, 501-550, 551-600 or 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referred to. The term "positional segments of EST-related nucleic acids" also includes segments comprising nucleotides 1-100, 101-200, 201-300, 301-400, 501-500, 500-600, or 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referred to. In addition, the term "positional segments of EST-related nucleic acids" includes segments comprising nucleotides 1-200, 201-400, 400-600, or 601-the terminal nucleotide of the EST-related nucleic acids to

the extent that such nucleotide positions are consistent with the lengths of the particular EST related nucleic acids being referred to. The present invention also includes the sequences complementary to the positional segments of EST-related nucleic acids.

[0033] The present invention also includes isolated, purified, or enriched "fragments of positional segments of EST-related nucleic acids." The terms "isolated", "purified", or "enriched" have the meanings provided above. As used herein, the term "fragments of positional segments of EST-related nucleic acids" refers to fragments comprising at least 10, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 150, or 200 consecutive nucleotides of the positional segments of EST-related nucleic acids. The present invention also includes the sequences complementary to the fragments of positional segments of EST-related nucleic acids .

[0034] The present invention also includes isolated or purified "EST-related polypeptides." The terms "isolated" or "purified" have the meanings provided above. As used herein, the term "EST-related polypeptides" means the polypeptides encoded by the EST-related nucleic acids, including the polypeptides of SEQ ID NOs: 4101-8177.

[0035] The present invention also includes isolated or purified "fragments of EST-related polypeptides." The terms "isolated" or "purified" have the meanings provided above. As used herein, the term "fragments of EST-related polypeptides" means fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of an EST-related polypeptide to the extent that fragments of these lengths are consistent with the lengths of the particular EST-related polypeptides being referred to.

[0036] The present invention also includes isolated or purified "positional segments of EST-related polypeptides." As used herein, the term "positional segments of EST-related polypeptides" includes polypeptides comprising amino acid residues 1-25, 26-50, 51-75, 76-100, 101-125, 126-150, 151-175, 176-200, or 201-the C-terminal amino acid of the EST-related polypeptides to the extent that such amino acid residues are consistent with the lengths of the particular EST-related polypeptides being referred to. The term "positional segments of EST-related polypeptides also includes segments comprising amino acid residues 1-50, 51-100, 101-150, 151-200 or 201-the C-terminal amino acid of the EST-related polypeptides to the extent that such amino acid residues are consistent with the lengths of the particular EST-related polypeptides being referred to. The term "positional segments of EST-related polypeptides" also includes segments comprising amino acids 1-100 or 101-200 of the EST-related polypeptides to the extent that such amino acid residues are consistent with the lengths of particular EST-related polypeptides being referred to. In addition, the term "positional segments of EST-related polypeptides" includes segments comprising amino acid residues 1-200 or 201-the C-terminal amino acid of the EST-related polypeptides to the extent that amino acid residues are consistent with the lengths of the particular EST related polypeptides being referred to.

[0037] The present invention also includes isolated or purified "fragments of positional segments of EST-related polypeptides." The terms "isolated" or "purified" have the meanings provided above. As used herein, the term "fragments of positional segments of EST-related polypeptides" means fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of positional segments of EST-related polypeptides to the extent that fragments of these lengths are consistent with the lengths of the particular EST-related polypeptides being referred to.

[0038] The present invention also includes antibodies which specifically recognize the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides. In the case of secreted proteins, such as those of SEQ ID NOs: 7798-7888 antibodies which specifically recognize the mature protein generated when the signal peptide is cleaved may also be obtained as described below. Similarly, antibodies which specifically recognize the signal peptides of SEQ ID NOs: 4101-4729 or 7798-7888 may also be obtained.

[0039] In some embodiments and in the case of secreted proteins, the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids include a signal sequence. In other embodiments, the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may include the full coding sequence for the protein or, in the case of secreted proteins, the full coding sequence of the mature protein (i.e. the protein generated when the signal polypeptide is cleaved off). In addition, the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may include regulatory regions upstream of the translation start site or downstream of the stop codon which control the amount, location, or developmental stage of gene expression.

[0040] As discussed above, both secreted and non-secreted human proteins may be therapeutically important. Thus, the proteins expressed from the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may be useful in treating or controlling a variety of human conditions.

[0041] The EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may be used in forensic procedures to identify individuals or in diagnostic procedures to identify individuals having genetic diseases resulting from abnormal gene expression. In addition, the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids are useful for constructing a high resolution map of the human chromosomes.

[0042] The present invention also relates to secretion vectors capable of directing the secretion of a protein of interest. Such vectors may be used in gene therapy strategies in which it is desired to produce a gene product in one cell which is to be delivered to another location in the body. Secretion vectors may also facilitate the

purification of desired proteins.

[0043] The present invention also relates to expression vectors capable of directing the expression of an inserted gene in a desired spatial or temporal manner or at a desired level. Such vectors may include sequences upstream of the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids, such as promoters or upstream regulatory sequences.

[0044] The present invention also comprises fusion vectors for making chimeric polypeptides comprising a first polypeptide and a second polypeptide. Such vectors are useful for determining the cellular localization of the chimeric polypeptides or for isolating, purifying or enriching the chimeric polypeptides.

[0045] The EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may also be used for gene therapy to control or treat genetic diseases. In the case of secreted proteins, signal peptides may be fused to heterologous proteins to direct their extracellular secretion.

[0046] Bacterial clones containing Bluescipt plasmids having inserts containing the sequence of the non-clustered 5'ESTs are presently stored at 80°C in 4% (v/v) glycerol in the inventor's laboratories under the designations. The non-clustered 5'ESTs are those which comprise a single EST from a single tissue in the listing of Table II. The inserts may be recovered from the stored materials by growing the appropriate clones on a suitable medium. The Bluescript DNA can then be isolated using plasmid isolation procedures familiar to those skilled in the art such as alkaline lysis minipreps or large scale alkaline lysis plasmid isolation procedures. If desired the plasmid DNA may be further enriched by centrifugation on a cesium chloride gradient, size exclusion chromatography, or anion exchange chromatography. The plasmid DNA obtained using these procedures may then be manipulated using standard cloning techniques familiar to those skilled in the art. Alternatively, a PCR can be done with primers designed at both ends of the inserted EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids. The PCR product which corresponds to the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids can then be manipulated using standard cloning techniques familiar to those skilled in the art.

[0047] One embodiment of the present invention is a purified nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

[0048] Another embodiment of the present invention is a purified nucleic acid comprising at least 10 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

[0049] Another embodiment of the present invention is a purified nucleic acid comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

[0050] A further embodiment of the present invention is a purified nucleic acid comprising the coding sequence of a sequence selected from the group consisting of 24-4100.

[0051] Yet another embodiment of the present invention is a purified nucleic acid comprising the full coding sequences of a sequence selected from the group consisting of SEQ ID NOs: 3721-3811 wherein the full coding sequence comprises the sequence encoding the signal peptide and the sequence encoding the mature protein.

Still another embodiment of the present invention is a purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs: 3721-3811 which encodes the mature protein.

[0052] Another embodiment of the present invention is a purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs: 24-652 and 3721-3811 which encodes the signal peptide.

[0053] Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

[0054] Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 7798-7888.

[0055] Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a mature protein included in a sequence selected from the group consisting of the sequences of SEQ ID NOs: 7798-7888.

[0056] Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a signal peptide included in a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-4729 and 7798-7888.

[0057] Another embodiment of the present invention is a purified nucleic acid at least 15,18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500 or 1000 nucleotides in length which hybridizes under stringent conditions to a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

[0058] Another embodiment of the present invention is a purified or isolated polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

[0059] Another embodiment of the present invention is a purified or isolated polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 7798-7888.

[0060] Another embodiment of the present invention is a purified or isolated polypeptide comprising a mature protein of a polypeptide selected from the group consisting of SEQ ID NOs: 7798-7888.

[0061] Another embodiment of the present invention is a purified or isolated polypeptide comprising a signal peptide of a sequence selected from the group consisting of the polypeptides of SEQ ID NOs: 4101-4729 and 7798-7888.

[0062] Another embodiment of the present invention is a purified or isolated polypeptide comprising at least 10 consecutive amino acids of a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

[0063] Another embodiment of the present invention is a method of making a cDNA comprising the steps of contacting a collection of mRNA molecules from human cells with a primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, hybridizing said primer to an mRNA in said collection that encodes said protein reverse transcribing said hybridized primer to make a first cDNA strand from said mRNA, making a second cDNA strand complementary to said first cDNA strand and isolating the resulting cDNA encoding said protein comprising said first cDNA strand and said second cDNA strand.

[0064] Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

[0065] In one aspect of this embodiment, the cDNA encodes at least a portion of a human polypeptide.

[0066] Another embodiment of the present invention is a method of making a cDNA comprising the steps of obtaining a cDNA comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, contacting said cDNA with a detectable probe comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 under conditions which permit said probe to hybridize to said cDNA, identifying a cDNA which hybridizes to said detectable probe, and isolating said cDNA which hybridizes to said probe.

[0067] Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

[0068] In one aspect of this embodiment, the cDNA encodes at least a portion of a human polypeptide.

[0069] Another embodiment of the present invention is a method of making a cDNA comprising the steps of contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of said mRNA, hybridizing said first primer to said polyA tail, reverse transcribing said mRNA to make a first cDNA strand, making a second cDNA strand complementary to said first cDNA strand using at least one primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, and isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

[0070] Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

[0071] In one aspect of this embodiment, said cDNA encodes at least a portion of a human polypeptide.

[0072] In another aspect of the preceding method the second cDNA strand is made by contacting said first cDNA strand with a first pair of primers, said first pair of primers comprising a second primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and a third primer having a sequence therein which is included within the sequence of said first primer, performing a first polymerase chain reaction with said first pair of primers to generate a first PCR product, contacting said first PCR product with a second pair of primers, said second pair of primers comprising a fourth primer, said fourth primer comprising at least 15 consecutive nucleotides of said sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, and a fifth primer, wherein said fourth and fifth hybridize to sequences within said first PCR product, and performing a second polymerase chain reaction, thereby generating a second PCR product.

[0073] One aspect of this embodiment is a purified cDNA obtainable by the method of the preceding paragraph.

[0074] In another aspect of this embodiment, said cDNA encodes at least a portion of a human polypeptide.

[0075] Alternatively, the second cDNA strand may be made by contacting said first cDNA strand with a second primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, hybridizing said second primer to said first strand cDNA, and extending said hybridized second primer to generate said second cDNA strand.

[0076] One aspect of the above embodiment is a purified cDNA obtainable by the method of the preceding paragraph.

[0077] In a further aspect of this embodiment said cDNA encodes at least a portion of a human polypeptide.

[0078] Another embodiment of the present invention is a method of making a polypeptide comprising the steps of obtaining a cDNA which encodes a polypeptide encoded by a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 or a cDNA which encodes a polypeptide comprising at least 10 consecutive amino acids of a polypeptide encoded by a sequence selected from the group consisting of SEQ ID NOs: 24-4100, inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter, introducing said expression vector into a host cell whereby said host cell produces the protein encoded by said cDNA, and isolating said protein.

[0079] Another aspect of this embodiment is an isolated protein obtainable by the method of the preceding paragraph.

[0080] Another embodiment of the present invention is a method of obtaining a promoter DNA comprising the steps of obtaining genomic DNA located upstream of a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, screening said genomic DNA to identify a promoter capable of directing transcription initiation, and
isolating said DNA comprising said identified promoter.

[0081] In one aspect of this embodiment, said obtaining step comprises walking from genomic DNA comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681. In another aspect of this embodiment, said screening step comprises inserting genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 into a promoter reporter vector. For example, said screening step may comprise identifying motifs in genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 which are transcription factor binding sites or transcription start sites.

[0082] Another embodiment of the present invention is a isolated promoter obtainable by the method of the paragraph above.
Another embodiment of the present invention is the inclusion of at least one sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and fragments comprising at least 15 consecutive nucleotides of said sequence in an array of discrete ESTs or fragments thereof of at least 15 nucleotides in length. In some aspects of this embodiment, the array includes at least two sequences selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, and fragments comprising at least 15 consecutive nucleotides of said sequences. In another aspect of this embodiment., the array includes at least five sequences selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and fragments comprising at least 15 consecutive nucleotides of said sequences.

[0083] Another embodiment of the present invention is an enriched population of recombinant nucleic acids, said recombinant nucleic acids comprising an insert nucleic acid and a backbone nucleic acid, wherein at least 5% of said insert nucleic acids in said population comprise a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

[0084] Another embodiment of the present invention is a purified or isolated antibody capable of specifically binding to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 4101-8177.
A purified or isolated antibody capable of specifically binding to a polypeptide comprising at least 10 consecutive amino acids of a sequence selected from the group consisting of SEQ ID NOs: 4101-8177.
An antibody composition capable of selectively binding to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 8 amino acids of any of SEQ ID NOs: 4101-8177, wherein said antibody is polyclonal or monoclonal.

[0085] Another embodiment of the present invention is a computer readable medium having stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177.

[0086] Another embodiment of the present invention is a computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177. In one aspect of this embodiment the computer system further comprises a sequence comparer and a data storage device having reference sequences stored thereon. For example, the sequence comparer may comprise a computer program which indicates polymorphisms.
In another aspect of this embodiment, the computer system further comprises an identifier which identifies features in said sequence.

[0087] Another embodiment of the present invention is a method for comparing a first sequence to a reference sequence wherein said first sequence is selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177 comprising the steps of reading said first sequence and said reference sequence through use of a computer program which compares sequences and determining differences between said first sequence and said reference sequence with said computer program. In some aspects of this embodiment, said step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

[0088] Another embodiment of the present invention is a method for identifying a feature in a sequence selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177 comprising the steps of reading said sequence through the use of a computer program which identifies features in sequences and identifying features in said sequence with said computer program.

[0089] Another embodiment of the present invention is a vector comprising a nucleic acid according to any one of

the nucleic acids described above.

[0090] Another embodiment of the present invention is a host cell containing the above vector.

[0091] Another embodiment of the present invention is a method of making any of the nucleic acids described above comprising the steps of introducing said nucleic acid into a host cell such that said nucleic acid is present in multiple copies in each host cell and isolating said nucleic acid from said host cell.

[0092] Another embodiment of the present invention is a method of making a nucleic acid of any of the nucleic acids described above comprising the step of sequentially linking together the nucleotides in said nucleic acids.

[0093] Another embodiment of the present invention is a method of making any of the polypeptides described above wherein said polypeptides is 150 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptide.

[0094] Another embodiment of the present invention is a method of making any of the polypeptides described above wherein said polypeptides is 120 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptides.

Brief Description of the Sequence Listing

[0095] SEQ ID NOs: 1, 3, 5, 7, 9, 11, and 13 are full-length cDNAs prepared using the methods described herein.

[0096] SEQ ID NOs: 2, 4, 6, 8, 10, 12, and 14 are the polypeptides encoded by the nucleic acids of SEQ ID NOs: 1, 3, 5, 7, 9, 11, and 13.

[0097] SEQ ID NOs: 15, 16, 18, 19, 21 and 22 are primers whose use is described in the specification.

[0098] SEQ ID NOs: 17, 20, and 23 are the sequences of nucleic acids containing transcription factor binding sites which were obtained as described below.

[0099] SEQ ID NOs: 24-652 are nucleic acids having an incomplete ORF which encodes a signal peptide. As used herein, an "incomplete ORF" is an open reading frame in which a start codon has been identified but no stop codon has been identified. The locations of the incomplete ORFs and sequences encoding signal peptides are listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein.

[0100] SEQ ID NOs: 653-3720 are nucleic acids having an incomplete ORF in which no sequence encoding a signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a sequence encoding a signal peptide in these nucleic acids. The locations of the incomplete ORFs are listed in the accompanying Sequence Listing.

[0101] SEQ ID NOs: 3721-3811 are nucleic acids having a complete ORF which encodes a signal peptide. As used herein, a "complete ORF" is an open reading frame in which a start codon and a stop codon have been identified. The locations of the complete ORFs and sequences encoding signal peptides are listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein.

[0102] SEQ ID NOs: 3812-4100 are nucleic acids having a complete ORF in which no sequence encoding a signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a sequence encoding a signal peptide in these nucleic acids. The locations of the complete ORFs are listed in the accompanying Sequence Listing.

[0103] SEQ ID NOs: 4101-4729 are "incomplete polypeptide sequences" which include a signal peptide. Incomplete polypeptide sequences" are polypeptide sequences encoded by nucleic acids in which a start codon has been identified but no stop codon has been identified. These polypeptides are encoded by the nucleic acids of SEQ ID NOs: 24-652. The location of the signal peptide is listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein.

[0104] SEQ ID NOs: 4730-7797 are incomplete polypeptide sequences in which no signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a signal peptide in these polypeptides. These polypeptides are encoded by the nucleic acids of SEQ ID NOs: 653-3720.

[0105] SEQ ID NOs: 7798-7888 are "complete polypeptide sequences" which include a signal peptide. "Complete polypeptide sequences" are polypeptide sequences encoded by nucleic acids in which a start codon and a stop codon have been identified. These polypeptides are encoded by the nucleic acids of SEQ ID NOs: 3721-3811. The location of the signal peptide is listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein.

[0106] SEQ ID NOs: 7889-8177 are complete polypeptide sequences in which no signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a signal peptide in these

polypeptides. These polypeptides are encoded by the nucleic acids of SEQ ID NOs.:3812-4100.

[0107] SEQ ID NOs: 8178-36681 are nucleic acid sequences in which no open reading frame has been conclusively identified to date. However, it remains possible subsequent analysis will identify an open reading frame in these nucleic acids.

[0108] In the accompanying Sequence Listing, all instances of the symbol "n" in the nucleic acid sequences mean that the nucleotide can be adenine, guanine, cytosine or thymine. In some instances the polypeptide sequences in the Sequence Listing contain the symbol "Xaa." These "Xaa" symbols indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined sequence where applicants believe one should not exist (if the sequence were determined more accurately). In some instances, several possible identities of the unknown amino acids may be suggested by the genetic code.

Brief Description of the Drawings

[0109] Figure 1 summarizes the computer analysis procedure for obtaining consensus contigated ESTs.

[0110] Figure 2 is an analysis of the 43 amino terminal amino acids of all human SwissProt proteins to determine the frequency of false positives and false negatives using the techniques for signal peptide identification described herein.

[0111] Figure 3 illustrates methods for making extended cDNAs.

[0112] Figure 4 provides a schematic description of the promoters isolated and the way they are assembled with the corresponding 5' tags.

[0113] Figure 5 describes the transcription factor binding sites present in each of these promoters.

Detailed Description of the Preferred Embodiment

I. General Methods for Obtaining 5' ESTs derived from mRNAs with intact 5' ends

[0114] In order to obtain the 5' ESTs of the present invention, mRNAs with intact 5' ends must be obtained. Example 1 below describes the preparation of 5' ESTs.

EXAMPLE 1

Preparation of mRNA

[0115] Total human RNAs or polyA$^+$à RNAs derived from 30 different tissues were respectively purchased from LABIMO and CLONTECH and used to generate 42 cDNA libraries as described below. The purchased RNA had been isolated from cells or tissues using acid guanidium thiocyanate-phenol-chloroform extraction (Chomczyniski and Sacchi, *Analytical Biochemistry* **162**:156-159, 1987). PolyA$^+$à RNA was isolated from total RNA (LABIMO) by two passes of oligo dT chromatography, as described by Aviv and Leder., *Proc. Natl. Acad. Sci. USA* 69:1408-1412, 1972) in order to eliminate ribosomal RNA.

[0116] The quality and the integrity of the polyA$^+$à RNAs were checked. Northern blots hybridized with a globin probe were used to confirm that the mRNAs were not degraded. Contamination of the polyA$^+$à mRNAs by ribosomal sequences was checked using Northern blots and a probe derived from the sequence of the 28S rRNA. Preparations of mRNAs with less than 5% of rRNAs were used in library construction. To avoid constructing libraries with RNAs contaminated by exogenous sequences (prokaryotic or fungal), the presence of bacterial 16S ribosomal sequences or of two highly expressed fungal mRNAs was examined using PCR.

[0117] Following preparation of the mRNAs from various tissues an oligonucleotide tag was specifically attached to the caps at the 5' ends of the mRNAs. The oligonucleotide tag had an EcoRI site therein to facilitate later cloning procedures. Following attachment of the oligonucleotide tag to the mRNA, the integrity of the mRNA was examined by performing a Northern blot with 200 to 500 ng of mRNA using a probe complementary to the oligonucleotide tag before performing the first strand synthesis described in Example 2.

EXAMPLE 2

cDNA Synthesis Using mRNA Templates Having Intact 5' Ends

[0118] For the mRNAs joined to oligonucleotide tags, first strand cDNA synthesis was performed using a reverse transcriptase with random nonamers as primers. In order to protect internal EcoRI sites in the cDNA from digestion at later steps in the procedure, methylated dCTP was used for first strand synthesis. After removal of RNA by an alkaline hydrolysis, the first strand of cDNA was precipitated using isopropanol in order to eliminate residual primers.

[0119] The second strand of the cDNA was synthesized with a Klenow fragment using a primer corresponding to the 5'end of the ligated oligonucleotide. Methylated dCTP was also used for second strand synthesis in order to protect internal EcoRI sites in the cDNA from digestion during the cloning process.

[0120] Following cDNA synthesis, the cDNAs were cloned into pBlueScript as described in Example 3 below.

EXAMPLE 3

Cloning of cDNAs derived from mRNA with intact 5' ends into BlueScript

[0121] Following second strand synthesis, the ends of the cDNA were blunted with T4 DNA polymerase (Biolabs) and the cDNA was digested with EcoRI. Since methylated dCTP was used during cDNA synthesis, the EcoRI site present in the tag was the only hemi-methylated site, hence the only site susceptible to EcoRI digestion. The cDNA was then size fractionated using exclusion chromatography (AcA, Biosepra) and fractions corresponding to cDNAs of more than 150 bp were pooled and ethanol precipitated. The cDNA was directionally cloned into the SmaI and EcoRI ends of the phagemid pBlueScript vector (Stratagene). The ligation mixture was electroporated into bacteria and propagated under appropriate antibiotic selection.

[0122] Clones containing the oligonucleotide tag attached were then selected as described in Example 4 below.

EXAMPLE 4

Selection of Clones Having the Oligonucleotide Tag Attached Thereto

[0123] The plasmid DNAs containing 5' EST libraries made as described above were purified (Qiagen). A positive selection of the tagged clones was performed as follows. Briefly, in this selection procedure, the plasmid DNA was converted to single stranded DNA using gene II endonuclease of the phage FI in combination with an exonuclease (Chang *et al.*, **Gene 127**:95-8, 1993) such as exonuclease III or T7 gene 6 exonuclease. The resulting single stranded DNA was then purified using paramagnetic beads as described by Fry *et al., Biotechniques,* **13**: 124-131, 1992. In this procedure, the single stranded DNA was hybridized with a biotinylated oligonucleotide having a sequence corresponding to the 3' end of the oligonucleotide tag. Clones including a sequence complementary to the biotinylated oligonucleotide were captured by incubation with streptavidin coated magnetic beads followed by magnetic selection. After capture of the positive clones, the plasmid DNA was released from the magnetic beads and converted into double stranded DNA using a DNA polymerase such as the Thermosequenase obtained from Amersham Pharmacia Biotech. The double stranded DNA was then electroporated into bacteria. The percentage of positive clones having the 5' tag oligonucleotide was estimated to typically rank between 90 and 98% using dot blot analysis.

[0124] Following electroporation, the libraries were ordered in 384-microtiter plates (MTP). A copy of the MTP was stored for future needs. Then the libraries were transferred into 96 MTP and sequenced as described below.

EXAMPLE 5

Sequencing of Inserts in Selected Clones

[0125] Plasmid inserts were first amplified by PCR on PE-9600 thermocyclers (Perkin-Elmer, Applied Biosystems Division, Foster City, CA), using standard SETA-A and SETA-B primers (Genset SA), AmpliTaqGold (Perkin-Elmer), dNTPs (Boehringer), buffer and cycling conditions as recommended by the Perkin-Elmer Corporation.

[0126] PCR products were then sequenced using automatic ABI Prism 377 sequencers (Perkin Elmer). Sequencing reactions were performed using PE 9600 thermocyclers with standard dye-primer chemistry and ThermoSequenase (Amersham Pharmacia Biotech). The primers used were either T7 or 21M13 (available from Genset SA) as appropriate. The primers were labeled with the JOE, FAM, ROX and TAMRA dyes. The dNTPs and ddNTPs used in the sequencing reactions were purchased from Boehringer. Sequencing buffer, reagent concentrations and cycling conditions were as recommended by Amersham.

[0127] Following the sequencing reaction, the samples were precipitated with ethanol, resuspended in formamide loading buffer, and loaded on a standard 4% acrylamide gel. Electrophoresis was performed for 2.5 hours at 3000V on an ABI 377 sequencer, and the sequence data were collected and analyzed using the ABI Prism DNA Sequencing Analysis Software, version 2.1.2.

EXAMPLE 6

Obtaining 5' ESTs from Full-length cDNA libraries Obtained from mRNA with Intact 5' Ends

[0128] Alternatively, 5'ESTs may be isolated from other cDNA or genomic DNA libraries. Such cDNA or genomic DNA libraries may be obtained from a commercial source or made using other techniques familiar to those skilled in the art. One example of such cDNA library construction, a full-length cDNA library, is as follows.

[0129] PolyA+ RNAs are prepared and their quality checked as described in Example 1. Then, the caps at the 5' ends of the polyA+ RNAs are specifically joined to an oligonucleotide tag. The oligonucleotide tag may contain a restriction site such as Eco RI to facilitate further subcloning procedures. Northern blotting is then performed to check the size of mRNAs having the oligonucleotide tag attached thereto and to ensure that the mRNAs were actually tagged.

[0130] First strand synthesis is subsequently carried out for mRNAs joined to the oligonucleotide tag as described in Example 2 above except that the random nonamers are replaced by an oligo-dT primer. For instance, this oligo-dT primer may contain an internal tag of 4 nucleotides which is different from one tissue to the other.

Following second strand synthesis using a primer contained in the oligonucleotide tag attached to the 5' end of mRNA, the blunt ends of the obtained double stranded full-length DNAs are modified into cohesive ends to facilitate subcloning. For example, the extremities of full-length cDNAs may be modified to allow subcloning into the Eco RI and Hind III sites of a Bluescript vector using the Eco RI site of the oligonucleotide tag and the addition of a Hind III adaptor to the 3' end of full-length cDNAs.

[0131] The full-length cDNAs are then separated into several fractions according to their sizes using techniques familiar to those skilled in the art. For example, electrophoretic separation may be applied in order to yield 3 or 6 different fractions. Following gel extraction and purification, the cDNA fractions are subcloned into appropriate vectors, such as Bluescript vectors, transformed into competent bacteria and propagated under appropriate antibiotic conditions. Subsequently, plasmids containing tagged full-length cDNAs are positively selected as described in Example 4.

[0132] The 5' end of full-length cDNAs isolated from such cDNA libraries may then be sequenced as described in Example 5

II.2. Computer Analysis of the Isolated 5' ESTs: construction of NetGene™ and SignalTag™ databases

[0133] The sequence data from the 42 cDNA libraries made as described above were transferred to a database, where quality control and validation steps were performed. A base-caller, working using a Unix system, automatically flagged suspect peaks, taking into account the shape of the peaks, the inter-peak resolution, and the noise level. The proprietary base-caller also performed an automatic trimming. Any stretch of 25 or fewer bases having more than 4 suspect peaks was considered unreliable and was discarded. Sequences corresponding to cloning vector or ligation oligonucleotides were automatically removed from the EST sequences. However, the resulting EST sequences may contain 1 to 5 bases belonging to the above mentioned sequences at their 5' end. If needed, these can easily be removed on a case to case basis.

[0134] Following sequencing as described above, the sequences of the 5' ESTs were entered in NetGene™, a database for storage and manipulation as described below and as depicted in Figure 1. Before searching the ESTs in the NetGene™ database for sequences of interest, ESTs derived from mRNAs which were not of interest, such as endogenous or exogenous contaminants, redundant sequences, small sequences, highly degenerate sequences, or repeated sequences were identified and eliminated from further consideration.

[0135] In order to determine the accuracy of the sequencing procedure as well as the efficiency of the 5' selection described above, the analyses described in Examples 7 and 8 respectively were performed on 5'ESTs obtained from NetGene™ database following the elimination of sequences which were not of interest.

EXAMPLE 7

Measurement of Sequencing Accuracy by Comparison to Known Sequences

[0136] To further determine the accuracy of the sequencing procedure described in Example 5, the sequences of NetGene™ 5' ESTs derived from known sequences were identified and compared to the original known sequences. First, a FASTA analysis with overhangs shorter than 5 bp on both ends was conducted on the 5' ESTs to identify those matching an entry in the public human mRNA database. The 6655 5' ESTs which matched a known human mRNA were then realigned with their cognate mRNA and dynamic programming was used to include substitutions, insertions, and deletions in the list of "errors" which would be recognized. Errors occurring in the last 10 bases of the 5' EST sequences were ignored to avoid the inclusion of spurious cloning sites in the analysis of sequencing accuracy.

[0137] This analysis revealed that the sequences incorporated in the NETGENE™ database had an accuracy of more than 99.5%.

EXAMPLE 8

Determination of Efficiency of 5' EST Selection

[0138] To determine the efficiency at which the above selection procedures isolated 5' ESTs which included sequences close to the 5' end of the mRNAs from which they derived, the sequences of the ends of the 5' ESTs derived from the elongation factor 1 subunit a and ferritin heavy chain genes were compared to the known cDNA sequences of these genes. Since the transcription start sites of both genes are well characterized, they may be used to determine the percentage of derived 5' ESTs which included the authentic transcription start sites.

[0139] For both genes, more than 95% of the obtained 5' ESTs actually included sequences close to or upstream of the 5' end of the corresponding mRNAs.

[0140] To extend the analysis of the reliability of the procedures for isolating 5' ESTs from ESTs in the NetGene™ database, a similar analysis was conducted using a database composed of human mRNA sequences extracted from GenBank database release 97 for comparison. The 5' ends of more than 85% of 5' ESTs derived from mRNAs included in the GeneBank database were located close to the 5' ends of the known sequence. As some of the mRNA sequences available in the GenBank database are deduced from genomic sequences, a 5' end matching with these sequences will be counted as an internal match. Thus, the method used here underestimates the yield of ESTs including the authentic 5' ends of their corresponding mRNAs.

EXAMPLE 9

Clustering of the 5' ESTs

[0141] Since the cDNA libraries made above include multiple 5' ESTs derived from the same mRNA, overlapping 5'ESTs may be assembled into continuous sequences. The following method (see Figure 1) describes how to efficiently cluster 5'ESTs in order to yield not only consensus 5'EST sequences for mRNAs derived from different genes but also consensus 5'EST sequences for different mRNAs, so called variants, transcribed from the same gene such as alternatively spliced mRNAs. This clustering was performed on a set of NetGene™ 5'ESTs sequences following elimination of endogenous contaminants, elimination of uninformative sequences and masking of repeats.

[0142] The whole set of sequences was first partitioned into smaller sets, so-called clusters, containing sequences exhibiting perfect matches with each other on a given length. Such clusters contain 5'ESTs derived from a small number of different genes. Some 5'EST sequences were not clustered using this approach either because they were not homologous to any other sequence or because the homology was not properly detected. To overcome this problem, sequences not clustered, so called singletons, may be compared to the consensus contigated ESTs obtained later on and, if necessary, included in the appropriate clusters and used to compute other consensus contigated ESTs.

[0143] Thereafter, all variants of a given gene were identified in each cluster as follows. Overlapping sequences inside a given cluster were figured as oriented graphs where each sequence was a node and each overlap an edge. Then, the different genes contained within a single graph which were represented by different connex components were identified and isolated from each other. Subsequently, the different variants of a same gene were isolated using an algorithm based on the detection of forks within a connex component. If desired, the consensus contigated EST sequences may be verified by identifying clones in nucleic acid samples derived from biological tissues, such as cDNA libraries, which hybridize to the probes based on the sequences of the consensus contigated ESTs and sequencing them.

[0144] Overlapping 5'EST sequences belonging to the same variant as well as included 5'EST sequences belonging to the same cluster were then contigated and consensus contigated 5'EST sequences were generated for each variant. Some of the obtained consensus contigated 5'EST sequences were incomplete due to the fact that only included and overlapping 5'EST sequences were considered to isolate genes and due to the algorithm developed to find variants. These variant consensus contigated 5'EST sequences were extended as follows. Variants transcribed from the same gene were compared pairwise and the 5' EST consensus sequences that were incomplete either in 5' and/or in 3' were extended with the appropriate sequence from the other variants. All 5' EST consensus sequences eventually completed in 5' or 3' from each cluster were subsequently compared to the whole set of individual 5'EST sequences obtained for this cluster.

EXAMPLE 10

Identification of the Most Probable Open Reading Frame of 5' ESTs

[0145] Subsequently, the most probable coding open reading frame (ORF) may be determined for each consensus assembled 5'EST or 5'EST as follows.

[0146] Each nucleic acid sequence is first divided into several subsequences which coding propensity is evaluated using different methods known to those skilled in the art such as the evaluation of N-mer frequency and its variants (Fickett and Tung, *Nucleic Acids* Res;**20**:6441-50 (1992)) or the Average Mutual Information method (Grosse *et al,* International Conference on Intelligent Systems for Molecular Biology, Montreal, Canada. June 28-July 1, 1998). Each of the scores obtained by the techniques described above are then normalized by their distribution extremities and then fused using a neural network into a unique score that represents the coding probability of a given subsequence.

[0147] The coding probability scores obtained for each subsequence, thus the probability score profiles obtained for each reading frame, are then linked to the initiation codons present on the sequence. For each open reading frame, defined as a nucleic acid sequence of at least 50 nucleotides beginning with an ATG codon, an ORF score is determined. Basically, this score is the sum of the probability scores computed for each subsequence corresponding to the considered ORF in the correct reading frame corrected by a function that negatively ponderates locally high score values and positively ponderates sustained high score values. The chosen ORF is the one with the highest score.

[0148] Two kinds of ORFs are considered. In some embodiments, 5'ESTs encoding ORFs of at least 50 amino acids extending up to the end of the consensus assembled 5'EST sequences are obtained. In other embodiments, 5'ESTs encoding complete ORFs, namely ORFs with start and stop codons, containing at least 100 amino acids are obtained.

EXAMPLE 11

Sequence Analysis

[0149] Application of the clustering method described in Example 9 to a selected set of 126,735 NetGene™ 5'ESTs free from endogenous contaminants and uninformative sequences yielded 9490 consensus assembled 5'EST

sequences or variants for a total of 8037 genes clustered representing 98,973 individual 5'ESTs. One of them which contained 21,138 sequences and was shown to contain chimeras thanks to comparison to public sequences was removed from further analysis.

[0150] Both non clustered 5'ESTs, *i.e.* singletons, and consensus contigated 5'ESTs were then compared to already known sequences as follows. Those sequences matching human mRNA sequences were eliminated from further analysis. Then, following masking of repeats those sequences matching sequences that have already been discovered by the inventors, namely sequences exhibiting more than 90% homology over stretches longer than 40 nucleotides using BLAST2N with overhangs shorter than 10 nucleotides, were removed from further consideration. The final set represents the sequences of the invention (SEQ ID NOs:24-4100 and 8178-36681),*i.e.*, 7609 consensus contigated 5'EST from 6398 clusters containing 31,267 5'ESTs and 24, 972 singletons.

[0151] Of the 6398 obtained clusters, 658 were shown to be multivariant, *i.e.* to contain several variants of the same gene. Table I gives for each of the multivariant clusters named by its internal reference (first column), the list of the consensus sequences of all variants, each variant being represented by a different SEQ ID NO.

[0152] Subsequently, the most probable open reading frame was determined, as described in Example 10, for all sequences of the invention. 3,697 5'ESTs (SEQ ID NOs:24-3720) encoding incomplete ORFs (SEQ ID NOs:4101-7797) of at least 50 amino acid long were found. In addition, 380 5'ESTs (SEQ ID NOs:3721-4100) encoding complete ORFs (SEQ ID NOs:7798-8177) of at least 100 amino acids were found.

[0153] The nucleotide sequences of the SEQ ID NOs: 24-4100 and 8178-36681 and the amino acid sequences encoded by SEQ ID NOs: 24-4100 (i.e. amino acid sequences of SEQ ID NOs: 4101-8177) are provided in the appended sequence listing. Some of the amino acid sequences may contain "Xaa" designators. These "Xaa" designators indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined sequence where applicants believe one should not exist (if the sequence were determined more accurately).

[0154] If one of the nucleic acid sequences of SEQ ID NOs: 24-4100 and 8178-36681 are suspected of containing one or more incorrect or ambiguous nucleotides, the ambiguities can readily be resolved by resequencing a fragment containing the nucleotides to be evaluated. If one or more incorrect or ambiguous nucleotides are detected, the corrected sequences should be included in the clusters from which the sequences were isolated, and used to compute other consensus contigated sequences on which other ORFs would be identified. Nucleic acid fragments for resolving sequencing errors or ambiguities may be obtained from deposited clones or can be isolated using the techniques described herein. Resolution of any such ambiguities or errors may be facilitated by using primers which hybridize to sequences located close to the ambiguous or erroneous sequences. For example, the primers may hybridize to sequences within 50-75 bases of the ambiguity or error. Upon resolution of an error or ambiguity, the corresponding corrections can be made in the protein sequences encoded by the DNA containing the error or ambiguity. The amino acid sequence of the protein encoded by a particular clone can also be determined by expression of the clone in a suitable host cell, collecting the protein, and determining its sequence.

[0155] In addition, if one of the sequences of SEQ ID NOs: 4101-8177 is suspected of containing an truncated ORF as the result of a frameshift in the sequence, such frameshifting errors may be corrected by combining the following two approaches. The first one involves thorough examination of all double predictions, i.e. all cases where the probability scores for two ORFs located on different reading frames are high and close, preferably different by less than 0.4. The fine examination of the region where the two possible ORFs overlap may help to detect the frameshift. In the second approach homologies with known proteins are used to correct suspected frameshifts.

EXAMPLE 12

Identification of Potential Signal Sequences in 5' ESTs

[0156] The amino acid sequences of SEQ ID NOs: 4101-8177 were then searched to identify potential signal motifs using slight modifications of the procedures disclosed in Von Heijne, *Nucleic Acids Res. 14*:4683-4690, 1986. Those sequences encoding a 15 amino acid long stretch with a score of at least 3.5 in the Von Heijne signal peptide identification matrix were considered to possess a signal sequence and were included in a database called SIGNALTAG™.

[0157] The sequences of the 720 nucleic acid sequences containing a signal sequence (SEQ ID NOs:24-652 and 3721-3811) and the corresponding polypeptides with a potential signal peptide (SEQ ID NO:4101-4729 and 7798-7888) are provided in the Sequence Listing appended hereto. The signal peptides of such polypeptides are indicated as features in the appended Sequence Listing. It should be noted that, in accordance with the regulations governing Sequence Listings, in the appended Sequence Listing, the full protein (i.e. the protein containing the signal peptide and the mature protein) extends from an amino acid residue having a negative number through a positively numbered C-terminal amino acid residue. Thus, the first amino acid of the mature protein resulting from cleavage of the signal peptide is designated as amino acid number 1, and the first amino acid of the signal peptide is designated with the appropriate negative number.

[0158] To confirm the accuracy of the above method for identifying signal sequences, the analysis of Example 13 was performed.

EXAMPLE 13

Confirmation of Accuracy of Identification of Potential Signal Sequences in 5' ESTs

[0159] The accuracy of the above procedure for identifying signal sequences encoding signal peptides was evaluated by applying the method to the 43 amino acids located at the N terminus of all human SwissProt proteins. The computed Von Heijne score for each protein was compared with the known characterization of the protein as being a secreted protein or a non-secreted protein. In this manner, the number of non-secreted proteins having a score higher than 3.5 (false positives) and the number of secreted proteins having a score lower than 3.5 (false negatives) could be calculated.

[0160] Using the results of the above analysis, the probability that a peptide encoded by the 5' region of the mRNA is in fact a genuine signal peptide based on its Von Heijne's score was calculated based on either the assumption that 10% of human proteins are secreted or the assumption that 20% of human proteins are secreted. The results of this analysis are shown in Figure 2.

[0161] Using the above method of identification of secretory proteins, 5' ESTs of the following polypeptides known to be secreted were obtained: human glucagon, gamma interferon induced monokine precursor, secreted cyclophilin-like protein, human pleiotropin, and human biotinidase precursor. Thus, the above method successfully identified those 5' ESTs which encode a signal peptide.

[0162] To confirm that the signal peptide encoded by the 5' ESTs or contigated consensus 5' ESTs actually functions as a signal peptide, the signal sequences from the 5' ESTs or consensus 5' ESTs may be cloned into a vector designed for the identification of signal peptides. Such vectors are designed to confer the ability to grow in selective medium only to host cells containing a vector with an operably linked signal sequence. For example, to confirm that a 5' EST or consensus 5' EST encodes a genuine signal peptide, the signal sequence of the 5' EST or consensus 5' EST may be inserted upstream and in frame with a non-secreted form of the yeast invertase gene in signal peptide selection vectors such as those described in U.S. Patent No. 5,536,637. Growth of host cells containing signal sequence selection vectors with the correctly inserted 5' EST or consensus 5' EST signal sequence confirms that the 5' EST or consensus 5' ESTs encodes a genuine signal peptide.

[0163] Alternatively, the presence of a signal peptide may be confirmed by cloning the extended cDNAs obtained using the ESTs or consensus 5' ESTs into expression vectors such as pXT1 as described below, or by constructing promoter-signal sequence-reporter gene vectors which encode fusion proteins between the signal peptide and an assayable reporter protein. After introduction of these vectors into a suitable host cell, such as COS cells or NIH 3T3 cells, the growth medium may be harvested and analyzed for the presence of the secreted protein. The medium from these cells is compared to the medium from control cells containing vectors lacking the signal sequence or extended cDNA insert to identify vectors which encode a functional signal peptide or an authentic secreted protein.

EXAMPLE 14

Assessment of the novelty rate of 5'ESTs

[0164] To assess the yield of new sequences, the obtained 5'ESTs and consensus contigated 5'ESTs were compared to all known human mRNAs extracted from the EMBL release 57 and daily updates available at the time of filing. The comparison was performed using BLAST2N on both strands following masking of the repeats. Sequences having more than 95% homology with public sequences over their whole length with at most 10 nucleotide overhangs on each extremity were considered as previously identified. Thus, about 90% of 5'ESTs or consensus assembled 5'ESTs were considered unidentified.

II. 3. Evaluation of Spatial and Temporal Expression of mRNAs Corresponding to the 5'ESTs or Extended cDNAs

[0165] Each of the SEQ ID NOs: 24-4100 and 8178-36681 was also categorized based on the tissue from which its corresponding mRNA was obtained, as described below in Example 15.

EXAMPLE 15

Expression Patterns of mRNAs From Which the 5'ESTs were obtained

[0166] Table II shows the spatial distribution of each of the 5'ESTs (non-clustered ESTs) and of each consensus contigated ESTs respectively. Table II provides the SEQ ID NOs: of the 5' ESTs (referred to alternatively herein as non-clustered ESTs or singletons) and consensus contigated ESTs. Table II also lists the number of ESTs from each type of tissue which were used to assemble the contigated consensus ESTs. The SEQ ID NOs: in Table II which contain a single 5' EST from a single tissue are 5' ESTs. Each type of tissue listed in Table II is encoded by a letter. The correspondence between the letter code and the tissue type is given in Table III. For example, the consensus contigated EST of SEQ ID NO: 47 contains one 5'EST from cancerous prostate, two 5'ESTs from lymph ganglia, and two 5'ESTs from testes.

[0167] In addition to categorizing the 5' ESTs and consensus contigated 5' ESTs with respect to their tissue of origin, the spatial and temporal expression patterns of the mRNAs corresponding to the 5' ESTs and consensus

contigated 5' ESTs, as well as their expression levels, may be determined as described in Example 16 below.

[0168] Characterization of the spatial and temporal expression patterns and expression levels of these mRNAs is useful for constructing expression vectors capable of producing a desired level of gene product in a desired spatial or temporal manner, as will be discussed in more detail below.

[0169] Furthermore, 5' ESTs and consensus contigated 5' ESTs whose corresponding mRNAs are associated with disease states may also be identified. For example, a particular disease may result from the lack of expression, over expression, or under expression of a mRNA corresponding to a 5' EST or consensus contigated 5' EST. By comparing mRNA expression patterns and quantities in samples taken from healthy individuals with those from individuals suffering from a particular disease, 5' ESTs or consensus contigated 5' ESTs responsible for the disease may be identified.

[0170] It will be appreciated that the results of the above characterization procedures for 5' ESTs and consensus contigated 5' ESTs also apply to extended cDNAs (obtainable as described below) which contain sequences adjacent to the 5' ESTs and consensus contigated 5' ESTs. It will also be appreciated that if desired, characterization may be delayed until extended cDNAs have been obtained rather than characterizing the 5' ESTs or consensus contigated 5' ESTs themselves.

EXAMPLE 16

Evaluation of Expression Levels and Patterns of mRNAs Corresponding to EST-Related Nucleic Acids

[0171] Expression levels and patterns of mRNAs corresponding to EST-related nucleic acids may be analyzed by solution hybridization with long probes as described in International Patent Application No. WO 97/05277. Briefly, an EST-related nucleic acid, fragment of an EST related nucleic acid, positional segment of an EST-related nucleic acid, or fragment of a positional segment of an EST-related nucleic acid corresponding to the gene encoding the mRNA to be characterized is inserted at a cloning site immediately downstream of a bacteriophage (T3, T7 or SP6) RNA polymerase promoter to produce antisense RNA. Preferably, the EST-related nucleic acid, fragment of an EST related nucleic acid, positional segment of an EST-related nucleic acid, or fragment of a positional segment of an EST-related nucleic acid is 100 or more nucleotides in length. The plasmid is linearized and transcribed in the presence of ribonucleotides comprising modified ribonucleotides (i.e. biotin-UTP and DIG-UTP). An excess of this doubly labeled RNA is hybridized in solution with mRNA isolated from cells or tissues of interest. The hybridizations are performed under standard stringent conditions (40-50°C for 16 hours in an 80% formamide, 0.4 M NaCl buffer, pH 7-8). The unhybridized probe is removed by digestion with ribonucleases specific for single-stranded RNA (i.e. RNases CL3, T1, Phy M, U2 or A). The presence of the biotin-UTP modification enables capture of the hybrid on a microtitration plate coated with streptavidin. The presence of the DIG modification enables the hybrid to be detected and quantified by ELISA using an anti-DIG antibody coupled to alkaline phosphatase.

[0172] The EST-related nucleic acid, fragment of an EST related nucleic acid, positional segment of an EST-related nucleic acid, or fragment of a positional segment of an EST-related nucleic acid may also be tagged with nucleotide sequences for the serial analysis of gene expression (SAGE) as disclosed in UK Patent Application No. 2 305 241 A. In this method, cDNAs are prepared from a cell, tissue, organism or other source of nucleic acid for which gene expression patterns must be determined. The resulting cDNAs are separated into two pools. The cDNAs in each pool are cleaved with a first restriction endonuclease, called an anchoring enzyme, having a recognition site which is likely to be present at least once in most cDNAs. The fragments which contain the 5' or 3' most region of the cleaved cDNA are isolated by binding to a capture medium such as streptavidin coated beads. A first oligonucleotide linker having a first sequence for hybridization of an amplification primer and an internal restriction site for a so called tagging endonuclease is ligated to the digested cDNAs in the first pool. Digestion with the second endonuclease produces short tag fragments from the cDNAs.

[0173] A second oligonucleotide having a second sequence for hybridization of an amplification primer and an internal restriction site is ligated to the digested cDNAs in the second pool. The cDNA fragments in the second pool are also digested with the tagging endonuclease to generate short tag fragments derived from the cDNAs in the second pool. The tags resulting from digestion of the first and second pools with the anchoring enzyme and the tagging endonuclease are ligated to one another to produce so called ditags. In some embodiments, the ditags are concatamerized to produce ligation products containing from 2 to 200 ditags. The tag sequences are then determined and compared to the sequences of the EST-related nucleic acid, fragment of an EST related nucleic acid, positional segment of an EST-related nucleic acid, or fragment of a positional segment of an EST-related nucleic acid to determine which 5' ESTs, contigated consensus 5' ESTs, or extended cDNAs are expressed in the cell, tissue, organism, or other source of nucleic acids from which the tags were derived. In this way, the expression pattern of the 5' ESTs, contigated consensus 5' ESTs, or extended cDNAs in the cell, tissue, organism, or other source of nucleic acids is obtained.

[0174] Quantitative analysis of gene expression may also be performed using arrays. As used herein, the term array means a one dimensional, two dimensional, or multidimensional arrangement of EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids. Preferably, the EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids are at least 15 nucleotides in length. More preferably, the EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids are at least 100 nucleotide long. More preferably, the fragments are more than 100 nucleotides

in length. In some embodiments, the EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids may be more than 500 nucleotides long.

[0175] For example, quantitative analysis of gene expression may be performed with EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids in a complementary DNA microarray as described by Schena *et al.* (*Science* **270**:467-470, 1995; *Proc. Natl. Acad. Sci. U.S.A.* 93:10614-10619, 1996). EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids are amplified by PCR and arrayed from 96-well microtiter plates onto silylated microscope slides using high-speed robotics. Printed arrays are incubated in a humid chamber to allow rehydration of the array elements and rinsed, once in 0.2% SDS for 1 min, twice in water for 1 min and once for 5 min in sodium borohydride solution. The arrays are submerged in water for 2 min at 95°C, transferred into 0.2% SDS for 1 min, rinsed twice with water, air dried and stored in the dark at 25°C.

[0176] Cell or tissue mRNA is isolated or commercially obtained and probes are prepared by a single round of reverse transcription. Probes are hybridized to 1 cm^2à microarrays under a 14 x 14 mm glass coverslip for 6-12 hours at 60°C. Arrays are washed for 5 min at 25°C in low stringency wash buffer (1 x SSC/0.2% SDS), then for 10 min at room temperature in high stringency wash buffer (0.1 x SSC/0.2% SDS). Arrays are scanned in 0.1 x SSC using a fluorescence laser scanning device fitted with a custom filter set. Accurate differential expression measurements are obtained by taking the average of the ratios of two independent hybridizations.

[0177] Quantitative analysis of the expression of genes may also be performed with EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids in complementary DNA arrays as described by Pietu *et al.* (*Genome Research* 6:492-503, 1996). The EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids thereof are PCR amplified and spotted on membranes. Then, mRNAs originating from various tissues or cells are labeled with radioactive nucleotides. After hybridization and washing in controlled conditions, the hybridized mRNAs are detected by phospho-imaging or autoradiography. Duplicate experiments are performed and a quantitative analysis of differentially expressed mRNAs is then performed.

[0178] Alternatively, expression analysis of the EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids can be done through high density nucleotide arrays as described by Lockhart *et al.* (*Nature Biotechnology* **14**: 1675-1680, 1996) and Sosnowsky *et al.* (*Proc. Natl. Acad. Sci.* **94**:1119-1123, 1997). Oligonucleotides of 15-50 nucleotides corresponding to sequences of EST-related nucleic acids, fragments of EST related nucleic acids, positional segments EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids are synthesized directly on the chip (Lockhart *et al., supra*) or synthesized and then addressed to the chip (Sosnowsky *et al., supra*). Preferably, the oligonucleotides are about 20 nucleotides in length.

[0179] cDNA probes labeled with an appropriate compound, such as biotin, digoxigenin or fluorescent dye, are synthesized from the appropriate mRNA population and then randomly fragmented to an average size of 50 to 100 nucleotides. The said probes are then hybridized to the chip. After washing as described in Lockhart *et al. supra* and application of different electric fields (Sonowsky et *al, supra.*), the dyes or labeling compounds are detected and quantified. Duplicate hybridizations are performed. Comparative analysis of the intensity of the signal originating from cDNA probes on the same target oligonucleotide in different cDNA samples indicates a differential expression of the mRNA corresponding to the 5' EST, consensus contigated 5' EST or extended cDNA from which the oligonucleotide sequence has been designed.

III. Use of 5' ESTs to Clone Extended cDNAs and to Clone the Corresponding Genomic DNAs

[0180] Once 5' ESTs or consensus contigated 5' ESTs which include the 5' end of the corresponding mRNAs have been selected using the procedures described above, they can be utilized to isolate extended cDNAs which contain sequences adjacent to the 5' ESTs or contigated consensus 5' ESTs. The extended cDNAs may include the entire coding sequence of the protein encoded by the corresponding mRNA, including the authentic translation start site. If the extended cDNA encodes a secreted protein, it may contain the signal sequence, and the sequence encoding the mature protein remaining after cleavage of the signal peptide. Extended cDNAs which include the entire coding sequence of the protein encoded by the corresponding mRNA are referred to herein as "full-length cDNAs." Alternatively, the extended cDNAs may not include the entire coding sequence of the protein encoded by the corresponding mRNA, although they do include sequences adjacent to the 5'ESTs or contigated consensus 5' ESTs. In some embodiments in which the extended cDNAs are derived from an mRNA encoding a secreted protein, the extended cDNAs may include only the sequence encoding the mature protein remaining after cleavage of the signal peptide, or only the sequence encoding the signal peptide.

[0181] Example 17 below describes a general method for obtaining extended cDNAs using 5' ESTs or consensus contigated 5' ESTs. Example 28 below describes the cloning and sequencing of several extended cDNAs, including extended cDNAs which include the entire coding sequence and authentic 5' end of the corresponding mRNA for several secreted proteins.

[0182] The methods of Examples 17 and 18 can also be used to obtain extended cDNAs which encode less than the entire coding sequence of proteins encoded by the genes corresponding to the 5' ESTs or consensus contigated

ESTs. In some embodiments, the extended cDNAs isolated using these methods encode at least 5,10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the proteins encoded by the sequences of SEQ ID NOs: 24-4100 and 8178-36681. In some embodiments, the extended cDNAs isolated using these methods encode at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the proteins encoded by the sequences of SEQ ID NOs: 24-4100.

EXAMPLE 17

General Method for Using 5' ESTs to Clone and Sequence Extended cDNAs which Include the Entire Coding Region and the Authentic 5'End of the Corresponding mRNA

[0183] The following general method has been used to quickly and efficiently isolate extended cDNAs including sequence adjacent to the sequences of the 5' ESTs used to obtain them. This method may be applied to obtain extended cDNAs for any 5' EST or consensus contigated 5' EST of the invention, including those 5' ESTs and consensus contigated 5' ESTs encoding secreted proteins. This method is summarized in Figure 3.

1. Obtaining Extended cDNAs

a) First strand synthesis

[0184] The method takes advantage of the known 5' sequence of the mRNA. A reverse transcription reaction is conducted on purified mRNA with a poly dT primer containing a nucleotide sequence at its 5' end allowing the addition of a known sequence at the end of the cDNA which corresponds to the 3' end of the mRNA. Such a primer and a commercially-available reverse transcriptase enzyme are added to a buffered mRNA sample yielding a reverse transcript anchored at the 3' polyA site of the RNAs. Nucleotide monomers are then added to complete the first strand synthesis.

[0185] After removal of the mRNA hybridized to the first cDNA strand by alkaline hydrolysis, the products of the alkaline hydrolysis and the residual poly dT primer can be eliminated with an exclusion column.

b) Second strand synthesis

[0186] A pair of nested primers on each end is designed based on the known 5' sequence from the 5' EST or contigated consensus 5' EST and the known 3' end added by the poly dT primer used in the first strand synthesis. Software used to design primers are either based on GC content and melting temperatures of oligonucleotides, such as OSP (Illier and Green, *PCR Meth. Appl.* 1:124-128, 1991), or based on the octamer frequency disparity method (Griffais *et al., Nucleic Acids Res.* 19: 3887-3891, 1991 such as PC-Rare (http: //bioinformatics.weizmann.ac.il/software/PC-Rare/doc/manuel.html).

[0187] Preferably, the nested primers at the 5' end and the nested primers at the 3' end are separated from one another by four to nine bases. These primer sequences may be selected to have melting temperatures and specificities suitable for use in PCR.

[0188] A first PCR run is performed using the outer primer from each of the nested pairs. A second PCR run is performed using the same enzyme and the inner primer from each of the nested pairs is then performed on a small sample of the first PCR product. Thereafter, the primers and remaining nucleotide monomers are removed.

2. Sequencing of Full Length Extended cDNAs or Fragments Thereof

[0189] Due to the lack of position constraints on the design of 5' nested primers compatible for PCR use using the OSP software, amplicons of two types are obtained. Preferably, the second 5' primer is located upstream of the translation initiation codon thus yielding a nested PCR product containing the entire coding sequence. Such a full length extended cDNA may be used in a direct cloning procedure. However, in some cases, the second 5' primer is located downstream of the translation initiation codon, thereby yielding a PCR product containing only part of the ORF. Such incomplete PCR products are submitted to a modified procedure described in section b below.

a) Nested PCR products containing complete ORFs

[0190] When the resulting nested PCR product contains the complete coding sequence, as predicted from the 5'EST or consensus contigated 5' EST sequence, it is cloned in an appropriate vector.

b) Nested PCR products containing incomplete ORFs

[0191] When the amplicon does not contain the complete coding sequence, intermediate steps are necessary to obtain both the complete coding sequence and a PCR product containing the full coding sequence. The complete coding sequence can be assembled from several partial sequences determined directly from different PCR products.

[0192] Once the full coding sequence has been completely determined, new primers compatible for PCR use are then designed to obtain amplicons containing the whole coding region. However, in such cases, 3' primers compatible for PCR use are located inside the 3' UTR of the corresponding mRNA, thus yielding amplicons which lack part of this

region, i.e. the polyA tract and sometimes the polyadenylation signal, as illustrated in Figure 3. Such full length extended cDNAs are then cloned into an appropriate vector.

c) Sequencing extended cDNAs

[0193] Sequencing of extended cDNAs can be performed using a Die Terminator approach with the AmpliTaq DNA polymerase FS kit available from Perkin Elmer.

[0194] In order to sequence PCR fragments, primer walking is performed using software such as OSP to choose primers and automated computer software such as ASMG (Sutton *et al., Genome Science Technol.* 1: 9-19, 1995) to construct contigs of walking sequences including the initial 5' tag using minimum overlaps of 32 nucleotides. Preferably, primer walking is performed until the sequences of full length cDNAs are obtained.

3. Cloning of Full Length Extended cDNAs

[0195] The PCR product containing the full coding sequence is then cloned in an appropriate vector. For example, the extended cDNAs can be cloned into any expression vector known in the art.

[0196] Since the PCR products obtained as described above are blunt ended molecules that can be cloned in either direction, the orientation of several clones for each PCR product is determined. Then, 4 to 10 clones are ordered in microtiter plates and subjected to a PCR reaction using a first primer located in the vector close to the cloning site and a second primer located in the portion of the extended cDNA corresponding to the 3' end of the mRNA. This second primer may be the antisense primer used in anchored PCR in the case of direct cloning (case a) or the antisense primer located inside the 3'UTR in the case of indirect cloning (case b). Clones in which the start codon of the extended cDNA is operably linked to the promoter in the vector so as to permit expression of the protein encoded by the extended cDNA are conserved and sequenced. In addition to the ends of cDNA inserts, approximately 50 bp of vector DNA on each side of the cDNA insert are also sequenced.

[0197] Cloned PCR products are then entirely sequenced in order to obtain at least two sequences per clone. Preferably, the sequences are obtained from both sense and antisense strands according to the aforementioned procedure with the following modifications. First, both 5' and 3' ends of cloned PCR products are sequenced in order to confirm the identity of the clone. Second, primer walking is performed if the full coding coding region has not been obtained yet. Contigation is then performed using primer walking sequences for cloned products as well as walking sequences that have already contigated for uncloned PCR products. The sequence is considered complete when the resulting contigs include the whole coding region as well as overlapping sequences with vector DNA on both ends. All the contigated sequences for each cloned amplicon are then used to obtain a consensus sequence.

4. Selection of cloned full length sequences obtained from the 5' ESTs of the present invention

[0198] A negative selection may be performed in order to eliminate unwanted cloned sequences resulting from either contaminants or PCR artifacts as follows. Sequences matching contaminant sequences such as vector DNA, tRNA, mtRNA, rRNA sequences are discarded as well as those encoding ORF sequences exhibiting extensive homology to repeats. Sequences obtained by direct cloning using nested primers on 5' and 3' tags (section 1. case a) but lacking polyA tail may be discarded. Only ORFs containing a signal peptide and ending either before the polyA tail (case a) or before the end of the cloned 3'UTR (case b) may be selected. Then, ORFs containing unlikely mature proteins such as mature proteins which size is less than 20 amino acids or less than 25% of the immature protein size may be eliminated.

[0199] Then, for each remaining full length extended cDNA containing several ORFs, a preselection of ORFs may be performed using the following criteria. The longest ORF with a signal peptide is preferred. If the ORF sizes are similar, the chosen ORF is the one which signal peptide has the highest score according to Von Heijne method

[0200] Sequences of full length extended cDNA clones may then be compared pairwise with BLAST after masking of the repeat sequences. Sequences containing at least 90% homology over 30 nucleotides may be clustered in the same class. Each cluster may then be subjected to a cluster analysis that detects sequences resulting from internal priming or from alternative splicing, identical sequences or sequences with several frameshifts. This automatic analysis serves as a basis for manual selection of the sequences.

[0201] Manual selection can be carried out using automatically generated reports for each sequenced full length extended cDNA clone. During this manual procedure, a selection is operated between clones belonging to the same class as follows.

[0202] Selection of full length extended cDNA clones encoding sequences of interest is performed using the following criteria. Structural parameters (initial tag, polyadenylation site and signal) may be checked. Then, homologies with known nucleic acids and proteins may be examined in order to determine whether the clone sequence match a known nucleic acid/protein sequence and, in the latter case, its covering rate and the date at which the sequence became public. Sequences resulting from chimera or double inserts or located on chromosome breaking points as assessed by homology to other sequences may be discarded during this procedure as well.

[0203] Extended cDNAs prepared as described above may be subsequently engineered to obtain nucleic acids which include desired portions of the extended cDNA using conventional techniques such as subcloning, PCR, or *in vitro* oligonucleotide synthesis. For example, if the extended cDNA is derived from a gene encoding a secreted polypeptide, it may include the full coding sequences (i.e. the sequences encoding the signal peptide and the mature

protein remaining after the signal peptide is cleaved off), the sequences encoding the mature polypeptide (i.e. the polypeptide generated after the signal peptide is cleaved off), or only the coding sequences for the signal peptides.

[0204] Similarly, nucleic acids containing any other desired portion of the coding sequences for the encoded protein may be obtained. For example, the nucleic acid may contain at least 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 consecutive bases of an extended cDNA.

[0205] Once an extended cDNA has been obtained, it can be sequenced to determine the amino acid sequence it encodes. Once the encoded amino acid sequence has been determined, one can create and identify any of the many conceivable cDNAs that will encode that protein by simply using the degeneracy of the genetic code. For example, allelic variants or other homologous nucleic acids can be identified as described below. Alternatively, nucleic acids encoding the desired amino acid sequence can be synthesized *in vitro*.

[0206] In a preferred embodiment, the coding sequence may be selected using the known codon or codon pair preferences for the host organism in which the cDNA is to be expressed.

[0207] In addition to PCR based methods for obtaining cDNAs which include the authentic 5'end of the corresponding mRNA as well as the full protein coding sequence of the corresponding mRNA, traditional hybridization based methods may also be employed. These methods may also be used to obtain the genomic DNAs which encode the mRNAs from which the 5' ESTs or contigated consensus 5' ESTs were derived, mRNAs corresponding to the extended cDNAs, or nucleic acids which are homologous to extended cDNAs, 5' ESTs, or contigated consensus 5' ESTs. Example 18 below provides examples of such methods.

[0208] Each identified ORF may be scanned for the presence of a signal peptide in the first 50 amino-acids or, where appropriate, within shorter regions down to 20 amino acids or less in the ORF, using the matrix method of von Heijne (*Nuc. Acids Res.* **14**: 4683-4690 (1986)) and the modification described in Example 12.

d) Homology to either nucleotide or protein sequences

[0209] Sequences of full-length extended cDNAs are then compared to known nucleotide sequences. Polypeptides encoded by full-length extended cDNAs are then compared to known polypeptide sequences.

[0210] Sequences of full-length extended cDNAs are compared to known nucleic acid sequences such as the vertebrate and EST sequences of Genbank, EMBL databases and Genseq (Derwent's database of patented nucleotide sequences). Full-length cDNA sequences are also compared to the sequences of a private database (Genset internal sequences) in order to find sequences that have already been identified by applicants. Sequences of full-length extended cDNAs with more than 90% homology over 30 nucleotides using either BLASTN or BLAST2N are identified as sequences that have already been described. Matching vertebrate sequences are subsequently examined using FASTA; full-length extended cDNAs with more than 70% homology over 30 nucleotides are identified as sequences that have already been described.

[0211] ORFs encoded by full-length extended cDNAs as defined in section c) are subsequently compared to known amino acid sequences found in public databases such as Swissprot, PIR and Genptept (Derwent's database of patented protein sequences). These analyses were performed using BLASTP with the parameter W=8 and allowing a maximum of 10 matches. Sequences of full-length extended cDNAs showing extensive homology to known protein sequences are recognized as already identified proteins.

[0212] In addition, the three-frame conceptual translation products of the top strand of full-length extended cDNAs are compared to publicly known amino acid sequences of Swissprot using BLASTX with the parameter E=0.001. Sequences of full-length extended cDNAs with more than 70% homology over 30 amino acid stretches are detected as already identified proteins.

5. Selection of cloned full-length sequences obtained from the 5' ESTs of the present invention

[0213] Cloned full-length extended cDNA sequences that have already been characterized by the aforementioned computer analysis are then submitted to an automatic procedure in order to preselect full-length extended cDNAs containing sequences of interest.

a) Automatic sequence preselection

[0214] All complete cloned full-length extended cDNAs clipped for vector on both ends are considered. First, a negative selection is operated in order to eliminate unwanted cloned sequences resulting from either contaminants or PCR artifacts as follows. Sequences matching contaminant sequences such as vector DNA; tRNA, mtRNA, rRNA sequences are discarded as well as those encoding ORF sequences exhibiting extensive homology to repeats as defined in section 4 a). Sequences obtained by direct cloning using nested primers on 5' and 3' tags (section 1. case a) but lacking polyA tail are discarded. Only ORFs containing a signal peptide and ending either before the polyA tail (case a) or before the end of the cloned 3'UTR (case b) are kept. Then, ORFs containing unlikely mature proteins such as mature proteins which size is less than 20 amino acids or less than 25% of the immature protein size are eliminated.

[0215] Then, for each remaining full-length extended cDNA containing several ORFs, a preselection of ORFs is performed using the following criteria. The longest ORF with a signal peptide is preferred. If the ORF sizes are similar, the chosen ORF is the one which signal peptide has the highest score according to Von Heijne method

[0216] Sequences of full-length extended cDNA clones are then compared pairwise with BLAST after masking of

the repeat sequences. Sequences containing at least 90% homology over 30 nucleotides are clustered in the same class. Each cluster is then subjected to a cluster analysis that detects sequences resulting from internal priming or from alternative splicing, identical sequences or sequences with several frameshifts. This automatic analysis serves as a basis for manual selection of the sequences.

b) Manual sequence selection

[0217] Manual selection can be carried out using automatically generated reports for each sequenced full-length extended cDNA clone. During this manual procedure, a selection is operated between clones belonging to the same class as follows. ORF sequences encoded by clones belonging to the same class are aligned and compared. If the homology between nucleotide sequences of clones belonging to the same class is more than 90% over 30 nucleotide stretches or if the homology between amino acid sequences of clones belonging to the same class is more than 80% over 20 amino acid stretches, than the clones are considered as being identical. The chosen ORF is either the one exhibiting matches with known amino acid sequences or the best one according to the criteria mentioned in the automatic sequence preselection section. If the nucleotide and amino acid homologies are less than 90% and 80% respectively, the clones are said to encode distinct proteins which can be both selected if they contain sequences of interest.

[0218] Selection of full-length extended cDNA clones encoding sequences of interest is performed using the following criteria. Structural parameters (initial tag, polyadenylation site and signal) are first checked. Then, homologies with known nucleic acids and proteins are examined in order to determine whether the clone sequence match a known nucleotide/protein sequence and, in the latter case, its covering rate and the date at which the sequence became public. If there is no extensive match with sequences other than ESTs or genomic DNA, or if the clone sequence brings substantial new information, such as encoding a protein resulting from alternative splicing of an mRNA coding for an already known protein, the sequence is kept. Examples of such cloned full-length extended cDNAs containing sequences of interest are described in Example 18. Sequences resulting from chimera or double inserts or located on chromosome breaking points as assessed by homology to other sequences are discarded during this procedure.

[0219] Extended cDNAs prepared as described above may be subsequently engineered to obtain nucleic acids which include desired portions of the extended cDNA using conventional techniques such as subcloning, PCR, or *in vitro* oligonucleotide synthesis. For example, nucleic acids which include only the full coding sequences (i.e. the sequences encoding the signal peptide and the mature protein remaining after the signal peptide is cleaved off) may be obtained using techniques known to those skilled in the art. Alternatively, conventional techniques may be applied to obtain nucleic acids which contain only the coding sequences for the mature protein remaining after the signal peptide is cleaved off or nucleic acids which contain only the coding sequences for the signal peptides.

[0220] Similarly, nucleic acids containing any other desired portion of the coding sequences for the encoded protein may be obtained. For example, the nucleic acid may contain at least 10, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive bases of an extended cDNA.

[0221] Once an extended cDNA has been obtained, it can be sequenced to determine the amino acid sequence it encodes. Once the encoded amino acid sequence has been determined, one can create and identify any of the many conceivable cDNAs that will encode that protein by simply using the degeneracy of the genetic code. For example, allelic variants or other homologous nucleic acids can be identified as described below. Alternatively, nucleic acids encoding the desired amino acid sequence can be synthesized *in vitro*.

[0222] In a preferred embodiment, the coding sequence may be selected using the known codon or codon pair preferences for the host organism in which the cDNA is to be expressed.

[0223] In addition to PCR based methods for obtaining cDNAs which include the authentic 5'end of the corresponding mRNA as well as the complete protein coding sequence of the corresponding mRNA, traditional hybridization based methods may also be employed. These methods may also be used to obtain the genomic DNAs which encode the mRNAs from which the 5' ESTs or consensus contigated 5' ESTS were derived, mRNAs corresponding to the extended cDNAs, or nucleic acids which are homologous to extended cDNAs, 5' ESTs, or consensus contigated 5' ESTs. Example 18 below provides examples of such methods.

EXAMPLE 18

Methods for Obtaining Extended cDNAs which Include the Entire Coding Region and the Authentic 5'End of the Corresponding mRNA or Nucleic Acids Homologous to Extended cDNAs, 5' ESTs or Consensus Contigated 5' ESTs

[0224] A full-length cDNA library can be made using the strategies described in Examples 1-4 above by replacing the random nonamer used in Example 2 with an oligo-dT primer. Alternatively, a cDNA library or genomic DNA library may be obtained from a commercial source or made using techniques familiar to those skilled in the art.

[0225] Such cDNA or genomic DNA libraries may be used to isolate extended cDNAs obtained from 5' ESTs or consensus contigated 5' ESTs or nucleic acids homologous to extended cDNAs, 5' ESTs, or consensus contigated 5' ESTs as follows. The cDNA library or genomic DNA library is hybridized to a detectable probe. The detectable probe may comprise at least 10, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive nucleotides of the 5' EST, consensus contigated 5' EST, or extended cDNA.

[0226] Techniques for identifying cDNA clones in a cDNA library which hybridize to a given probe sequence are

disclosed in Sambrook *et al., Molecular Cloning: A Laboratory Manual 2d Ed.*, Cold Spring Harbor Laboratory Press, 1989. The same techniques may be used to isolate genomic DNAs.

[0227] Briefly, cDNA or genomic DNA clones which hybridize to the detectable probe are identified and isolated for further manipulation as follows. The detectable probe described in the preceding paragraph is labeled with a detectable label such as a radioisotope or a fluorescent molecule. Techniques for labeling the probe are well known and include phosphorylation with polynucleotide kinase, nick translation, *in vitro* transcription, and non radioactive techniques. The cDNAs or genomic DNAs in the library are transferred to a nitrocellulose or nylon filter and denatured. After blocking of non specific sites, the filter is incubated with the labeled probe for an amount of time sufficient to allow binding of the probe to cDNAs or genomic DNAs containing a sequence capable of hybridizing thereto.

[0228] By varying the stringency of the hybridization conditions used to identify cDNAs or genomic DNAs which hybridize to the detectable probe, cDNAs or genomic DNAs having different levels of homology to the probe can be identified and isolated as described below.

1. Identification of cDNA or Genomic DNA Sequences Having a High Degree of Homology to the Labeled Probe

[0229] To identify cDNAs or genomic DNAs having a high degree of homology to the probe sequence, the melting temperature of the probe may be calculated using the following formulas:

[0230] For probes between 14 and 70 nucleotides in length the melting temperature (Tm) is calculated using the formula: Tm=81.5+16.6(log [Na+])+0.41(fraction G+C)-(600/N) where N is the length of the probe.

[0231] If the hybridization is carried out in a solution containing formamide, the melting temperature may be calculated using the equation Tm=81.5+16.6(log [Na+])+0.41(fraction G+C)-(0.63% formamide)-(600/N) where N is the length of the probe.

[0232] Prehybridization may be carried out in 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 µg denatured fragmented salmon sperm DNA or 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 µg denatured fragmented salmon sperm DNA, 50% formamide. The formulas for SSC and Denhardt's solutions are listed in Sambrook*et al., supra.*

[0233] Hybridization is conducted by adding the detectable probe to the prehybridization solutions listed above. Where the probe comprises double stranded DNA, it is denatured before addition to the hybridization solution. The filter is contacted with the hybridization solution for a sufficient period of time to allow the probe to hybridize to extended cDNAs or genomic DNAs containing sequences complementary thereto or homologous thereto. For probes over 200 nucleotides in length, the hybridization may be carried out at 15-25°C below the Tm. For shorter probes, such as oligonucleotide probes, the hybridization may be conducted at 15-25°C below the Tm. Preferably, for hybridizations in 6X SSC, the hybridization is conducted at approximately 68°C. Preferably, for hybridizations in 50% formamide containing solutions, the hybridization is conducted at approximately 42°C.

[0234] All of the foregoing hybridizations would be considered to be under "stringent" conditions.

[0235] Following hybridization, the filter is washed in 2X SSC, 0.1% SDS at room temperature for 15 minutes. The filter is then washed with 0.1X SSC, 0.5% SDS at room temperature for 30 minutes to 1 hour. Thereafter, the solution is washed at the hybridization temperature in 0.1X SSC, 0.5% SDS. A final wash is conducted in 0.1X SSC at room temperature.

[0236] cDNAs or genomic DNAs which have hybridized to the probe are identified by autoradiography or other conventional techniques.

2. Obtaining cDNA or Genomic DNA Sequences Having Lower Degrees of Homology to the Labeled Probe

[0237] The above procedure may be modified to identify cDNAs or genomic DNAs having decreasing levels of homology to the probe sequence. For example, to obtain cDNAs or genomic DNAs of decreasing homology to the detectable probe, less stringent conditions may be used. For example, the hybridization temperature may be decreased in increments of 5°C from 68°C to 42°C in a hybridization buffer having a sodium concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate" conditions above 50°C and "low" conditions below 50°C.

[0238] Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42°C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of homology to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide.

[0239] cDNAs or genomic DNAs which have hybridized to the probe are identified by autoradiography.

3. Determination of the Degree of Homology between the Obtained cDNAs or Genomic DNAs and 5'ESTs, Consensus Contigated 5'ESTs, or Extended cDNAs or Between the Polypeptides Encoded by the Obtained cDNAs or Genomic DNAs and the Polypeptides Encoded by the 5'ESTs, Consensus Contigated 5'ESTs, or Extended cDNAs

[0240] To determine the level of homology between the hybridized cDNA or genomic DNA and the 5'EST, consensus contigated 5'EST or extended cDNA from which the probe was derived, the nucleotide sequences of the hybridized nucleic acid and the 5'EST, consensus contigated 5'EST or extended cDNA from which the probe was derived are compared. The sequences of the 5'EST, consensus contigated 5'EST or extended cDNA from which the

probe was derived and the sequences of the cDNA or genomic DNA which hybridized to the detectable probe may be stored on a computer readable medium as described below and compared to one another using any of a variety of algorithms familiar to those skilled in the art, those described below.

[0241] To determine the level of homology between the polypeptide encoded by the hybridizing cDNA or genomic DNA and the polypeptide encoded by the 5'EST, consensus contigated 5'EST or extended cDNA from which the probe was derived, the polypeptide sequence encoded by the hybridized nucleic acid and the polypeptide sequence encoded by the 5'EST, consensus contigated 5'EST or extended cDNA from which the probe was derived are compared. The sequences of the polypeptide encoded by the 5'EST, consensus contigated 5'EST or extended cDNA from which the probe was derived and the polypeptide sequence encoded by the cDNA or genomic DNA which hybridized to the detectable probe may be stored on a computer readable medium as described below and compared to one another using any of a variety of algorithms familiar to those skilled in the art, those described below.

[0242] Protein and/or nucleic acid sequence homologies may be evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, 1988, *Proc. Nail. Acad. Sci. USA 85(8)*:2444-2448; Altschul *et al.*, 1990,*J. Mol. Biol. 215(3)*:403-410; Thompson *et al.*, 1994, *Nucleic Acids Res. 22(2)*:4673-4680; Higgins *et al.*, 1996, *Methods Enzymol. 266*:383-402; Altschul*et al.*, 1990, *J. Mol. Biol. 215(3)*:403-410; Altschul *et al.*, 1993, *Nature Genetics* 3:266-272).

[0243] In a particularly preferred embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, *e.g.*, Karlin and Altschul, 1990,*Proc. Natl. Acad. Sci. USA 87*:2267-2268; Altschul *et al.*, 1990, *J. Mol. Biol. 215*:403-410; Altschul *et al.*, 1993, *Nature Genetics* 3:266-272; Altschul *et al.*, 1997, *Nuc. Acids Res. 25*:3389-3402). In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

[0244] The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (*i.e.*, aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet *et al.*, 1992, *Science 256*:1443-1445; Henikoff and Henikoff, 1993, *Proteins 17*:49-61). Less preferably, the PAM or PAM250 matrices may also be used (see, *e.g.*, Schwartz and Dayhoff, eds., 1978, *Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure.* Washington: National Biomedical Research Foundation)

[0245] The BLAST programs evaluate the statistical significance of all high-scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, e.g., Karlin and Altschul, 1990, *Proc. Natl. Acad. Sci. USA 87*:2267-2268).

[0246] The parameters used with the above algorithms may be adapted depending on the sequence length and degree of homology studied. In some embodiments, the parameters may be the default parameters used by the algorithms in the absence of instructions from the user.

[0247] In some embodiments, the level of homology between the hybridized nucleic acid and the extended cDNA, 5'EST, or 5' consensus contigated EST from which the probe was derived may be determined using the FASTDB algorithm described in Brutlag et al. Comp. App. Biosci. 6:237-245, 1990. In such analyses the parameters may be selected as follows: Matrix=Unitary, k-tuple=4, Mismatch Penalty=1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score=1, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the sequence which hybridizes to the probe, whichever is shorter. Because the FASTDB program does not consider 5' or 3' truncations when calculating homology levels, if the sequence which hybridizes to the probe is truncated relative to the sequence of the extended cDNA, 5'EST, or consensus contigated 5'EST from which the probe was derived the homology level is manually adjusted by calculating the number of nucleotides of the extended cDNA, 5'EST, or consensus contigated 5' EST which are not matched or aligned with the hybridizing sequence, determining the percentage of total nucleotides of the hybridizing sequence which the non-matched or non-aligned nucleotides represent, and subtracting this percentage from the homology level. For example, if the hybridizing sequence is 700

nucleotides in length and the extended cDNA, 5'EST, or consensus contigated 5' EST sequence is 1000 nucleotides in length wherein the first 300 bases at the 5' end of the extended cDNA, 5'EST, or consensus contigated 5' EST are absent from the hybridizing sequence, and wherein the overlapping 700 nucleotides are identical, the homology level would be adjusted as follows. The non-matched, non-aligned 300 bases represent 30% of the length of the extended cDNA, 5'EST, or consensus contigated 5' EST. If the overlapping 700 nucleotides are 100% identical, the adjusted homology level would be 100-30=70% homology. It should be noted that the preceding adjustments are only made when the non-matched or non-aligned nucleotides are at the 5' or 3' ends. No adjustments are made if the non-matched or non-aligned sequences are internal or under any other conditions.

[0248] For example, using the above methods, nucleic acids having at least 95% nucleic acid homology, at least 96% nucleic acid homology, at least 97% nucleic acid homology, at least 98% nucleic acid homology, at least 99% nucleic acid homology, or more than 99% nucleic acid homology to the extended cDNA, 5'EST, or consensus contigated 5' EST from which the probe was derived may be obtained and identified. Such nucleic acids may be allelic variants or related nucleic acids from other species. Similarly, by using progressively less stringent hybridization conditions one can obtain and identify nucleic acids having at least 90%, at least 85%, at least 80% or at least 75% homology to the extended cDNA, 5'EST, or consensus contigated 5' EST from which the probe was derived.

[0249] Using the above methods and algorithms such as FASTA with parameters depending on the sequence length and degree of homology studied, for example the default parameters used by the algorithms in the absence of instructions from the user, one can obtain nucleic acids encoding proteins having at least 99%, at least 98%, at least 97%, at least 96%, at least 95%, at least 90%, at least 85%, at least 80% or at least 75% homology to the protein encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST from which the probe was derived. In some embodiments, the homology levels can be determined using the "default" opening penalty and the "default" gap penalty, and a scoring matrix such as PAM 250 (a standard scoring matrix; see Dayhoff et al., in: Atlas of Protein Sequence and Structure, Vol. 5, Supp. 3 (1978)).

[0250] Alternatively, the level of polypeptide homology may be determined using the FASTDB algorithm described by Brutlag et al. Comp. App. Biosci. 6:237-245, 1990. In such analyses the parameters may be selected as follows: Matrix=PAM 0, k-tuple=2, Mismatch Penalty=1, Joining Penalty=20, Randomization Group Length=0, Cutoff Score=1, Window Size=Sequence Length, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the homologous sequence, whichever is shorter. If the homologous amino acid sequence is shorter than the amino acid sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST as a result of an N terminal and/or C terminal deletion the results may be manually corrected as follows. First, the number of amino acid residues of the amino acid sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST which are not matched or aligned with the homologous sequence is determined. Then, the percentage of the length of the sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST which the non-matched or non-aligned amino acids represent is calculated. This percentage is subtracted from the homology level. For example wherein the amino acid sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST is 100 amino acids in length and the length of the homologous sequence is 80 amino acids and wherein the amino acid sequence encoded by the extended cDNA or 5'EST is truncated at the N terminal end with respect to the homologous sequence, the homology level is calculated as follows. In the preceding scenario there are 20 non-matched, non-aligned amino acids in the sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST. This represents 20% of the length of the amino acid sequence encoded by the extended cDNA, 5'EST, or consensus contigated 5' EST. If the remaining amino acids are 1005 identical between the two sequences, the homology level would be 100%-20%=80% homology. No adjustments are made if the non-matched or non-aligned sequences are internal or under any other conditions.

[0251] In addition to the above described methods, other protocols are available to obtain extended cDNAs using 5' ESTs or consensus contigated 5'ESTs as outlined in the following paragraphs.

[0252] Extended cDNAs may be prepared by obtaining mRNA from the tissue, cell, or organism of interest using mRNA preparation procedures utilizing polyA selection procedures or other techniques known to those skilled in the art. A first primer capable of hybridizing to the polyA tail of the mRNA is hybridized to the mRNA and a reverse transcription reaction is performed to generate a first cDNA strand.

[0253] The first cDNA strand is hybridized to a second primer containing at least 10 consecutive nucleotides of the sequences of SEQ ID NOs 24-4100 and 8178-36681. Preferably, the primer comprises at least 10, 12, 15, 17, 18, 20, 23, 25, or 28 consecutive nucleotides from the sequences of SEQ ID NOs 24-4100 and 8178-36681. In some embodiments, the primer comprises more than 30 nucleotides from the sequences of SEQ ID NOs 24-4100 and 8178-36681. If it is desired to obtain extended cDNAs containing the full protein coding sequence, including the authentic translation initiation site, the second primer used contains sequences located upstream of the translation initiation site. The second primer is extended to generate a second cDNA strand complementary to the first cDNA strand. Alternatively, RT-PCR may be performed as described above using primers from both ends of the cDNA to be obtained.

[0254] Extended cDNAs containing 5' fragments of the mRNA may be prepared by hybridizing an mRNA comprising the sequences of SEQ ID NOs: 24-4100 and 8178-36681 with a primer comprising a complementary to a fragment of an EST-related nucleic acid hybridizing the primer to the mRNAs, and reverse transcribing the hybridized primer to make a first cDNA strand from the mRNAs. Preferably, the primer comprises at least 10, 12, 15, 17, 18, 20, 23, 25, or 28 consecutive nucleotides of the sequences complementary to SEQ ID NOs: 24-4100 and 8178-36681.

[0255] Thereafter, a second cDNA strand complementary to the first cDNA strand is synthesized. The second cDNA strand may be made by hybridizing a primer complementary to sequences in the first cDNA strand to the first cDNA strand and extending the primer to generate the second cDNA strand.

[0256] The double stranded extended cDNAs made using the methods described above are isolated and cloned. The extended cDNAs may be cloned into vectors such as plasmids or viral vectors capable of replicating in an appropriate host cell. For example, the host cell may be a bacterial, mammalian, avian, or insect cell.

[0257] Techniques for isolating mRNA, reverse transcribing a primer hybridized to mRNA to generate a first cDNA strand, extending a primer to make a second cDNA strand complementary to the first cDNA strand, isolating the double stranded cDNA and cloning the double stranded cDNA are well known to those skilled in the art and are described *in Current Protocols in Molecular Biology,* John Wiley & Sons, Inc. 1997 and Sambrook *et al.,* Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, 1989.

[0258] Alternatively, other procedures may be used for obtaining full-length cDNAs or extended cDNAs. In one approach, full-length or extended cDNAs are prepared from mRNA and cloned into double stranded phagemids as follows. The cDNA library in the double stranded phagemids is then rendered single stranded by treatment with an endonuclease, such as the Gene II product of the phage FI and an exonuclease (Chang *et al.,* **Gene 127**:95-8, 1993). A biotinylated oligonucleotide comprising the sequence of a fragment of an EST-related nucleic acid is hybridized to the single stranded phagemids. Preferably, the fragment comprises at least 10, 12, 15, 17, 18, 20, 23, 25, or 28 consecutive nucleotides of the sequences of SEQ ID NOs: 24-4100 and 8178-36681.

[0259] Hybrids between the biotinylated oligonucleotide and phagemids are isolated by incubating the hybrids with streptavidin coated paramagnetic beads and retrieving the beads with a magnet (Fry *et al., Biotechniques,* **13**: 124-131, 1992). Thereafter, the resulting phagemids are released from the beads and converted into double stranded DNA using a primer specific for the 5' EST or consensus contigated 5'EST sequence used to design the biotinylated oligonucleotide. Alternatively, protocols such as the Gene Trapper kit (Gibco BRL) may be used. The resulting double stranded DNA is transformed into bacteria. Extended cDNAs or full length cDNAs containing the 5' EST or consensus contigated 5'EST sequence are identified by colony PCR or colony hybridization.

[0260] Using any of the above described methods in section III, a plurality of extended cDNAs containing full-length protein coding sequences or portions of the protein coding sequences may be provided as cDNA libraries for subsequent evaluation of the encoded proteins or use in diagnostic assays as described below.

EXAMPLE 19

Full Length cDNAs

[0261] The procedures described in Example 17 and 18 were used to obtain 376 extended cDNAs or full length cDNAs derived from 5' ESTs in a variety of tissues. The following list provides a few examples of thus obtained cDNAs.

[0262] Using this procedure, the full length cDNA of SEQ ID NO:1 (internal identification number 58-34-2-E7-FL2) was obtained. This cDNA encodes the signal peptide MWWFQQGLSFLPSALVIWTSA (SEQ ID NO:2) having a von Heijne score of 5.5.

[0263] Using this approach, the full length cDNA of SEQ ID NO:3 (internal identification number 48-19-3-G1-FL1) was obtained. This cDNA encodes the signal peptide MKKVLLLITAILAVAVG (SEQ ID NO: 4) having a von Heijne score of 8.2.

[0264] The full length cDNA of SEQ ID NO:5 (internal identification number 58-35-2-F10-FL2) was also obtained using this procedure. This cDNA encodes a signal peptide LWLLFFLVTAIHA (SEQ ID NO:6) having a von Heijne score of 10.7.

[0265] Furthermore, the polypeptides encoded by the extended or full-length cDNAs may be screened for the presence of known structural or functional motifs or for the presence of signatures, small amino acid sequences which are well conserved amongst the members of a protein family. The results obtained for the polypeptides encoded by a few full-length cDNAs derived from 5'ESTs that were screened for the presence of known protein signatures and motifs using the Proscan software from the GCG package and the Prosite 15.0 database are provided below.

[0266] The protein of SEQ ID NO: 8 encoded by the full-length cDNA SEQ ID NO: 7 (internal designation 78-8-3-E6-CL0_1C) and expressed in adult prostate belong to the phosphatidylethanolamine-binding protein from which it exhibits the characteristic PROSITE signature from positions 90 to 112. Proteins from this widespread family, from nematodes to fly, yeast, rodent and primate species, bind hydrophobic ligands such as phospholipids and nucleotides. They are mostly expressed in brain and in testis and are thought to play a role in cell growth and/or maturation, in regulation of the sperm maturation, motility and in membrane remodeling. They may act either through signal transduction or through oxidoreduction reactions (for a review see Schoentgen and Jollès, *FEBS Letters,* **369** :22-26 (1995)). Taken together, these data suggest that the protein of SEQ ID NO: 8 may play a role in cell growth, maturation and in membrane remodeling and/or may be related to male fertility. Thus, these protein may be useful in diagnosing and/or treating cancer, neurodegenerative diseases, and/or disorders related to male fertility and sterility.

[0267] The protein of SEQ ID NO :10 encoded by the full-length cDNA SEQ ID NO:9 (internal designation 108-013-5-O-H9-FLC) shows homologies with a family of lysophospholipases conserved among eukaryotes (yeast, rabbit, rodents and human). In addition, some members of this family exhibit a calcium-independent phospholipase A2 activity (Portilla *et al, J. Am. Soc. Nephro.,*9 :1178-1186 (1998)). All members of this family exhibit the active site consensus GXSXG motif of carboxylesterases that is also found in the protein of SEQ ID NO :10 (position 54 to 58). In addition, this protein may be a membrane protein with one transmembrane domain as predicted by the software TopPred II (Claros and von Heijne, *CABIOS applic. Notes,* **10** :685-686 (1994)). Taken together, these data suggest that the protein of SEQ ID NO:10 may play a role in fatty acid metabolism, probably as a phospholipase. Thus, this protein or

part therein, may be useful in diagnosing and/or treating several disorders including, but not limited to, cancer, diabetes, and neurodegenerative disorders such as Parkinson's and Alzheimer's diseases. It may also be useful in modulating inflammatory responses to infectious agents and/or to suppress graft rejection.

[0268] The protein of SEQ ID NO: 12 encoded by the full-length cDNA SEQ ID NO: 11 (internal designation 108-004-5-0-D10-FLC) shows remote homology to a subfamily of beta4-galactosyltransferases widely conserved in animals (human, rodents, cow and chicken). Such enzymes, usually type II membrane proteins located in the endoplasmic reticulum or in the Golgi apparatus, catalyzes the biosynthesis of glycoproteins, glycolipid glycans and lactose. Their characteristic features defined as those of subfamily A in Breton *et al, J. Biochem.*, **123**:1000-1009 (1998) are pretty well conserved in the protein of SEQ ID NO: 12, especially the region I containing the DVD motif (positions 163-165) thought to be involved either in UDP binding or in the catalytic process itself. In addition, the protein of SEQ ID NO: 12 has the typical structure of a type II protein. Indeed, it contains a short 28-amino-acid-long N-terminal tail, a transmembrane segment from positions 29 to 49 and a large 278-amino-acid-long C-terminal tail as predicted by the software TopPred II (Claros and von Heijne, *CABIOS applic. Notes*, **10** :685-686 (1994)). Taken together, these data suggest that the protein of SEQ ID NO: 12 may play a role in the biosynthesis of polysaccharides, and of the carbohydrate moieties of glycoproteins and glycolipids and/or in cell-cell recognition. Thus, this protein may be useful in diagnosing and/or treating several types of disorders including, but not limited to, cancer, atherosclerosis, cardiovascular disorders, autoimmune disorders and rheumatic diseases including rheumatoid arthritis.

[0269] The protein of SEQ ID NO: 14 encoded by the full-length cDNA SEQ ID NO: 13 (internal designation 108-009-5-0-A2-FLC) shows extensive homology to the bZIP family of transcription factors, and especially to the human luman protein (Lu *et al., Mol. Cell. Biol.*, **17** :5117-5126 (1997))). The match include the whole bZIP domain composed of a basic DNA-binding domain and of a leucine zipper allowing protein dimerization. The basic domain is conserved in the protein of SEQ ID NO: 14 as shown by the characteristic PROSITE signature (positions 224-237) except for a conservative substitution of a glutamic acid with an aspartic acid in position 233. The typical PROSITE signature for leucine zipper is also present (positions 259 to 280). Taken together, these data suggest that the protein of SEQ ID NO: 14 may bind to DNA, hence regulating gene expression as a transcription factor. Thus, this protein may be useful in diagnosing and/or treating several types of disorders including, but not limited to, cancer.

[0270] Bacterial clones containing plasmids containing the full length cDNAs described above are presently stored in the inventor's laboratories under the internal identification numbers provided above. The inserts may be recovered from the deposited materials by growing an aliquot of the appropriate bacterial clone in the appropriate medium. The plasmid DNA can then be isolated using plasmid isolation procedures familiar to those skilled in the art such as alkaline lysis minipreps or large scale alkaline lysis plasmid isolation procedures. If desired the plasmid DNA may be further enriched by centrifugation on a cesium chloride gradient, size exclusion chromatography, or anion exchange chromatography. The plasmid DNA obtained using these procedures may then be manipulated using standard cloning techniques familiar to those skilled in the art. Alternatively, a PCR can be done with primers designed at both ends of the EST insertion. The PCR product which corresponds to the 5'EST can then be manipulated using standard cloning techniques familiar to those skilled in the art.

IV. Expression of Proteins

[0271] EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, and fragments of positional segments of EST-related nucleic acids may be used to express the polypeptides which they encode. In particular, they may be used to express EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides. In some embodiments, the EST-related nucleic acids, positional segments of EST-related nucleic acids, and fragments of positional segments of EST-related nucleic acids may be used to express the full polypeptide (i.e. the signal peptide and the mature polypeptide) of a secreted protein, the mature protein (i.e. the polypeptide generated after cleavage of the signal peptide), or the signal peptide of a secreted protein. If desired, nucleic acids encoding the signal peptide may be used to facilitate secretion of the expressed protein. It will be appreciated that a plurality of EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids may be simultaneously cloned into expression vectors to create an expression library for analysis of the encoded proteins as described below.

EXAMPLE 20

Expression of the Proteins Encoded by the Genes Corresponding to the 5'ESTs or Consensus Contigated 5' ESTs

[0272] To express their encoded proteins the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of EST-related nucleic acids are cloned into a suitable expression vector. In some instances, nucleic acids encoding EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides may be cloned into a suitable expression vector.

[0273] In some embodiments, the nucleic acids inserted into the expression vector may comprise the coding sequence of a sequence selected from the group consisting of 24-4100. In other embodiments, the nucleic acids inserted into the expression vector may comprise may comprise the full coding sequence (i.e. the nucleotides

encoding the signal peptide and the mature polypeptide) of one of SEQ ID NOs: 3721-3811. In some embodiments, the nucleic acid inserted into the expression vector may comprise the nucleotides of one of the sequences of SEQ ID NOs: 3721-3811 which encode the mature polypeptide (i.e. the nucleotides encoding the polypeptide generated after cleavage of the signal peptide). In further embodiments, the nucleic acids inserted into the expression vector may comprise the nucleotides of 24-652 and 3721-3811 which encode the signal peptide to facilitate secretion of the expressed protein. The nucleic acids inserted into the expression vectors may also contain sequences upstream of the sequences encoding the signal peptide, such as sequences which regulate expression levels or sequences which confer tissue specific expression.

[0274] The nucleic acid inserted into the expression vector may encode a polypeptide comprising the one of the sequences of SEQ ID NOs: 4101-8177. In some embodiments, the nucleic acid inserted into the expression vector may encode the full polypeptide sequence (i.e. the signal peptide and the mature polypeptide) included in one of SEQ ID NOs: 7798-7888. In other embodiments, the nucleic acid inserted into the expression vector may encode the mature polypeptide (i.e. the polypeptide generated after cleavage of the signal peptide) included in one of the sequences of SEQ ID NOs: 798-7888. In further embodiments, the nucleic acids inserted into the expression vector may encode the signal peptide included in one of the sequences of 4101-4729 and 7798-7888.

[0275] The nucleic acid encoding the protein or polypeptide to be expressed is operably linked to a promoter in an expression vector using conventional cloning technology. The expression vector may be any of the mammalian, yeast, insect or bacterial expression systems known in the art. Commercially available vectors and expression systems are available from a variety of suppliers including Genetics Institute (Cambridge, MA), Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence may be optimized for the particular expression organism in which the expression vector is introduced, as explained by Hatfield, *et al.*, U.S. Patent No. 5,082,767.

[0276] The following is provided as one exemplary method to express the proteins encoded by the nucleic acids described above. In some instances the nucleic acid encoding the protein or polypeptide to be expressed includes a methionine initiation codon and a polyA signal. If the nucleic acid encoding the polypeptide to be expressed lacks a methionine to serve as the initiation site, an initiating methionine can be introduced next to the first codon of the nucleic acid using conventional techniques. Similarly, if the nucleic acid encoding the protein or polypeptide to be expressed lacks a polyA signal, this sequence can be added to the construct by, for example, splicing out the polyA signal from pSG5 (Stratagene) using BglI and SalI restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene). pXT1 contains the LTRs and a portion of the*gag* gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex thymidine kinase promoter and the selectable neomycin gene. The nucleic acid encoding the polypeptide to be expressed is obtained by PCR from the bacterial vector using oligonucleotide primers complementary to the nucleic acid encoding the protein or polypeptide to be expressed and containing restriction endonuclease sequences for Pst I incorporated into the 5'primer and BglII at the 5' end of 3' primer, taking care to ensure that the nucleic acid encoding the protein or polypeptide to be expressed is correctly positioned with respect to the poly A signal. The purified fragment obtained from the resulting PCR reaction is digested with PstI, blunt ended with an exonuclease, digested with Bgl II, purified and ligated to pXT1, now containing a poly A signal and digested with BglII.

[0277] The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Technologies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600 µg/ml G418 (Sigma, St. Louis, Missouri).

[0278] Alternatively, the nucleic acid encoding the protein or polypeptide to be expressed may be cloned into pED6dpc2 as described above. The resulting pED6dpc2 constructs may be transfected into a suitable host cell, such as COS 1 cells. Methotrexate resistant cells are selected and expanded. The expressed protein or polypeptide may be isolated, purified, or enriched as described above.

[0279] To confirm expression of the desired protein or polypeptide, the proteins or polypeptides produced by cells containing a vector with a nucleic acid insert encoding the protein or polypeptide are compared to those lacking such an insert. The expressed proteins are detected using techniques familiar to those skilled in the art such as Coomassie blue or silver staining or using antibodies against the protein or polypeptide encoded by the nucleic acid insert. Antibodies capable of specifically recognizing the protein of interest may be generated using synthetic 15-mer peptides having a sequence encoded by the appropriate nucleic acid. The synthetic peptides are injected into mice to generate antibody to the polypeptide encoded by the nucleic acid.

[0280] If the proteins or polypeptides encoded by the nucleic acid inserts are secreted, medium prepared from the host cells or organisms containing an expression vector which contains a nucleic acid insert encoding the desired protein or polypeptide is compared to mdieum prepared from the control cells or organism. The presence of a band in medium from the cells containing the nucleic acid insert which is absent from preparations from the control cells indicates that the protein or polypeptide encoded by the nucleic acid insert is being expressed and secreted. Generally, the band corresponding to the protein encoded by the nucleic acid insert will have a mobility near that expected based on the number of amino acids in the open reading frame of the nucleic acid insert. However, the band may have a mobility different than that expected as a result of modifications such as glycosylation, ubiquitination, or enzymatic cleavage.

[0281] Alternatively, if the protein expressed from the above expression vectors does not contain sequences directing its secretion, the proteins expressed from host cells containing an expression vector with an insert

encoding a secreted protein or portion thereof can be compared to the proteins expressed in control host cells containing the expression vector without an insert. The presence of a band in samples from cells containing the expression vector with an insert which is absent in samples from cells containing the expression vector without an insert indicates that the desired protein or portion thereof is being expressed. Generally, the band will have the mobility expected for the secreted protein or portion thereof. However, the band may have a mobility different than that expected as a result of modifications such as glycosylation, ubiquitination, or enzymatic cleavage.

[0282] The expressed protein or polypeptide may be purified, isolated or enriched using a variety of methods. In some methods, the protein or polypeptide may be secreted into the culture medium via a native signal peptide or a heterologous signal peptide operably linked thereto. In some methods, the protein or polypeptide may be linked to a heterologous polypeptide which facilitates its isolation, purification, or enrichment such as a nickel binding polypeptide. The protein or polypeptide may also be obtained by gel electrophoresis, ion exchange chromatography, size chromatography, hplc, salt precipitation, immunoprecipitation, a combination of any of the preceding methods, or any of the isolation, purification, or enrichment techniques familiar to those skilled in the art.

[0283] The protein encoded by the nucleic acid insert may also be purified using standard immunochromatography techniques using immunoaffinity chromatography with antibodies directed against the encoded protein or polypeptide as described in more detail below. If antibody production is not possible, the nucleic acid insert encoding the desired protein or polypeptide may be incorporated into expression vectors designed for use in purification schemes employing chimeric polypeptides. In such strategies, the coding sequence of the nucleic acid insert is ligated in frame with the gene encoding the other half of the chimera. The other half of the chimera may be β-globin or a nickel binding polypeptide. A chromatography matrix having antibody to β-globin or nickel attached thereto is then used to purify the chimeric protein. Protease cleavage sites may be engineered between the β-globin gene or the nickel binding polypeptide and the extended cDNA or portion thereof. Thus, the two polypeptides of the chimera may be separated from one another by protease digestion.

[0284] One useful expression vector for generating β-globin chimerics is pSG5 (Stratagene), which encodes rabbit β-globin. Intron II of the rabbit β-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques as described are well known to those skilled in the art of molecular biology. Standard methods are published in methods texts such as Davis *et al.*, (*Basic Methods in Molecular Biology,* L.G. Davis, M.D. Dibner, and J.F. Battey, ed., Elsevier Press, NY, 1986) and many of the methods are available from Stratagene, Life Technologies, Inc., or Promega. Polypeptide may additionally be produced from the construct using *in vitro* translation systems such as the *In vitro* Express™ Translation Kit (Stratagene).

[0285] Following expression and purification of the proteins or polypeptides encoded by the nucleic acid inserts, the purified proteins may be tested for the ability to bind to the surface of various cell types as described in Example 21 below. It will be appreciated that a plurality of proteins expressed from these nucleic acid inserts may be included in a panel of proteins to be simultaneously evaluated for the activities specifically described below, as well as other biological roles for which assays for determining activity are available.

EXAMPLE 21

Analysis of Secreted Proteins to Determine Whether they Bind to the Cell Surface

[0286] The EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, fragments of positional segments of EST-related nucleic acids, nucleic acids encoding the EST-related polypeptides, nucleic acids encoding fragments of the EST-related polypeptides, nucleic acids encoding positional segments of EST-related polypeptides, or nucleic acids encoding fragments of positional segments of EST-related polypeptides are cloned into expression vectors such as those described in Example 20. The encoded proteins or polypeptides are purified, isolated, or enriched as described above. Following purification, isolation, or enrichment, the proteins or polypeptides are labeled using techniques known to those skilled in the art. The labeled proteins or polypeptides are incubated with cells or cell lines derived from a variety of organs or tissues to allow the proteins to bind to any receptor present on the cell surface. Following the incubation, the cells are washed to remove non-specifically bound proteins or polypeptides. The specifically bound labeled proteins or polypeptides are detected by autoradiography. Alternatively, unlabeled proteins or polypeptides may be incubated with the cells and detected with antibodies having a detectable label, such as a fluorescent molecule, attached thereto.

[0287] Specificity of cell surface binding may be analyzed by conducting a competition analysis in which various amounts of unlabeled protein or polypeptide are incubated along with the labeled protein or polypeptide. The amount of labeled protein or polypeptide bound to the cell surface decreases as the amount of competitive unlabeled protein or polypeptide increases. As a control, various amounts of an unlabeled protein or polypeptide unrelated to the labeled protein or polypeptide is included in some binding reactions. The amount of labeled protein or polypeptide bound to the cell surface does not decrease in binding reactions containing increasing amounts of unrelated unlabeled protein, indicating that the protein or polypeptide encoded by the nucleic acid binds specifically to the cell surface.

[0288] As discussed above, human proteins have been shown to have a number of important physiological effects and, consequently, represent a valuable therapeutic resource. The human proteins or polypeptides made as described above may be evaluated to determine their physiological activities as described below.

EXAMPLE 22

Assaying the Expressed Proteins or Polypeptides for Cytokine, Cell Proliferation or Cell Differentiation Activity

[0289] As discussed above, some human proteins act as cytokines or may affect cellular proliferation or differentiation. Many protein factors discovered to date, including all known cytokines, have exhibited activity in one or more factor dependent cell proliferation assays, and hence the assays serve as a convenient confirmation of cytokine activity. The activity of a protein or polypeptide of the present invention is evidenced by any one of a number of routine factor dependent cell proliferation assays for cell lines including, without limitation, 32D, DA2, DA1G, T10, B9, B9/11, BaF3, MC9/G, M+à (preB M+à), 2E8, RB5, DA1, 123, T1165, HT2, CTLL2, TF-1, Mo7c and CMK. The proteins or polypeptides prepared as described above may be evaluated for their ability to regulate T cell or thymocyte proliferation in assays such as those described above or in the following references:*Current Protocols in Immunology,* Ed. by J.E. Coligan *et al.,* Greene Publishing Associates and Wiley-Interscience; Takai *et al. J. Immunol.* **137**:3494-3500, 1986., Bertagnolli *et al.* J. Immunol. **145**:1706-1712, 1990., Bertagnolli *et al., Cellular Immunology* **133**:327-341, 1991. Bertagnolli, *et al.* ***J. Immunol.*** **149**:3778-3783, 1992; Bowman *et al., J. Immunol.* **152**:1756-1761, 1994.

[0290] In addition, numerous assays for cytokine production and/or the proliferation of spleen cells, lymph node cells and thymocytes are known. These include the techniques disclosed in **Current Protocols in Immunology**. J.E. Coligan *et al.* Eds., 1:3.12.1-3.12.14, John Wiley and Sons, Toronto. 1994; and Schreiber, R.D. In*Current Protocols in Immunology., supra* **1** : 6.8.1-6.8.8.

[0291] The proteins or polypeptides prepared as described above may also be assayed for the ability to regulate the proliferation and differentiation of hematopoietic or lymphopoietic cells. Many assays for such activity are familiar to those skilled in the art, including the assays in the following references: Bottomly *et al.,* In *Current Protocols in Immunology., supra.* 1 : 6.3.1-6.3.12,; deVries *et al., J.Exp. Med.* **173**:1205-1211, 1991; Moreau *et al., Nature* **36**:690-692, 1988; Greenberger *et al., Proc. Natl. Acad. Sci. U.S.A.* **80**:2931-2938, 1983; Nordan, R., In*Current Protocols in Immunology., supra.* **1** : 6.6.1-6.6.5; Smith et al., *Proc. Natl. Acad. Sci. U.S.A.* **83**:1857-1861, 1986; Bennett *et al* in *Current Protocols in Immunology supra* 1 : 6.15.1; Ciarletta *et al* In *Current Protocols in Immunology. supra* **1** : 6.13.1.

[0292] The proteins or polypeptides prepared as described above may also be assayed for their ability to regulate T-cell responses to antigens. Many assays for such activity are familiar to those skilled in the art, including the assays described in the following references: Chapter 3 (*In vitro* Assays for Mouse Lymphocyte Function), Chapter 6 (Cytokines and Their Cellular Receptors) and Chapter 7, (Immunologic Studies in Humans) in *Current Protocols in Immunology supra;* Weinberger *et al., Proc. Natl. Acad. Sci. USA* **77**:6091-6095, 1980; Weinberger *et al., Eur. J. Immun.* **11**:405-411, 1981; Takai *et al., J. Immunol.* **137**:3494-3500, 1986; Takai *et al., J. Immunol.* **140**:508-512, 1988.

[0293] Those proteins or polypeptides which exhibit cytokine, cell proliferation, or cell differentiation activity may then be formulated as pharmaceuticals and used to treat clinical conditions in which induction of cell proliferation or differentiation is beneficial. Alternatively, as described in more detail below, nucleic acids encoding these proteins or polypeptides or nucleic acids regulating the expression of these proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 23

Assaying the Expressed Proteins or Polypeptides for Activity as Immune System Regulators

[0294] The proteins or polypeptides prepared as described above may also be evaluated for their effects as immune regulators. For example, the proteins or polypeptides may be evaluated for their activity to influence thymocyte or splenocyte cytotoxicity. Numerous assays for such activity are familiar to those skilled in the art including the assays described in the following references: Chapter 3 (*In vitro* Assays for Mouse Lymphocyte Function 3.1-3.19) and Chapter 7 (Immunologic studies in Humans) in *Current Protocols in Immunology* , J.E. Coligan et al. Eds, Greene Publishing Associates and Wiley-Interscience; Herrmann *et al., Proc. Natl. Acad. Sci. USA* **78**:2488-2492, 1981; *Herrrnann et al., J Immunol.* **128**:1968-1974, 1982; Handa *et al., J. Immunol.* **135**:1564-1572, 1985; Takai *et al., J. Immunol.* **137**:3494-3500, 1986; Takai *et al., J. Immunol.* **140**:508-512, 1988; Bowman *et al., J. Virology***61**:1992-1998; Bertagnolli *et al. Cell. Immunol.* **133**:327-341, 1991; Brown *et al., J.Immunol.* **153**:3079-3092, 1994.

[0295] The proteins or polypeptides prepared as described above may also be evaluated for their effects on T-cell dependent immunoglobulin responses and isotype switching. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Maliszewski, *J. Immunol.* **144**:3028-3033, 1990; Mond *et al.* in *Current Protocols in Immunology,* 1 : 3.8.1-3.8.16, *supra.*

[0296] The proteins or polypeptides prepared as described above may also be evaluated for their effect on immune effector cells, including their effect on Th1 cells and cytotoxic lymphocytes. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Chapter 3 (*In vitro* Assays for Mouse Lymphocyte Function 3.1-3.19) and Chapter 7 (Immunologic Studies in Humans) in *Current Protocols in Immunology, supra;* Takai *et al., J. Immunol.* **137**:3494-3500, 1986; Takai *et al.; J. Immunol.* 140:508-512, 1988; Bertagnolli *et al., J. Immunol.* **149**:3778-3783, 1992.

[0297] The proteins or polypeptides prepared as described above may also be evaluated for their effect on

dendritic cell mediated activation of naive T-cells. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Guery *et al., J. Immunol.* **134**:536-544, 1995; *Inaba et al., J. Exp. Med.* **173**:549-559, 1991; Macatonia *et al., J. Immunol.* **154**:5071-5079, 1995; Porgador *et al J. Exp. Med* **182**:255-260, 1995; Nair*et al., J. Virol.* **67**:4062-4069, 1993; Huang *et al., Science* **264**:961-965, 1994; Macatonia *et al J. Exp. Med* **169**:1255-1264, 1989; Bhardwaj *et al., Journal of Clinical Investigation* **94**:797-807, 1994; and Inaba *et al., J. Exp. Med* **172**:631-640, 1990.

[0298] The proteins or polypeptides prepared as described above may also be evaluated for their influence on the lifetime of lymphocytes. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Darzynkiewicz *et al., Cytometry* **13**:795-808, 1992; Gorczyca *et al., Leukemia* **7**:659-670, 1993; Gorczyca *et al., Cancer Res.* **53**:1945-1951, 1993; Itoh *et al., Cell* **66**:233-243, 1991; Zacharchuk, *J. Immunol.* **145**:4037-4045, 1990; Zamai *et al., Cytometry* **14**:891-897, 1993; Gorczyca *et al., Int. J. Oncol.***1**:639-648, 1992.

[0299] The proteins or polypeptides prepared as described above may also be evaluated for their influence on early steps of T-cell commitment and development. Numerous assays for such activity are familiar to those skilled in the art, including without limitation the assays disclosed in the following references: Antica *et al., Blood* **84**:111-117, 1994; Fine *et al., Cell. Immunol.* **155**:111-122, 1994; Galy *et al., Blood* **85**:2770-2778, 1995; Toki *et al., Proc. Nat. Acad Sci. USA* **88**:7548-7551, 1991.

[0300] Those proteins or polypeptides which exhibit activity as immune system regulators activity may then be formulated as pharmaceuticals and used to treat clinical conditions in which regulation of immune activity is beneficial. For example, the protein or polypeptide may be useful in the treatment of various immune deficiencies and disorders (including severe combined immunodeficiency), e.g., in regulating (up or down) growth and proliferation of T and/or B lymphocytes, as well as effecting the cytolytic activity of NK cells and other cell populations. These immune deficiencies may be genetic or be caused by viral (e.g., HIV) as well as bacterial or fungal infections, or may result from autoimmune disorders. More specifically, infectious diseases caused by viral, bacterial, fungal or other infection may be treatable ucing the protein or polypeptide including infections by HIV, hepatitis viruses, herpesviruses, mycobacteria, Leishmania spp., plamodium. and various fungal infections such as candidiasis. Of course, in this regard, a protein or polypeptide may also be useful where a boost to the immune system generally may be desirable, i.e., in the treatment of cancer.

[0301] Alternatively, the proteins or polypeptides prepared as described above may be used in treatment of autoimmune disorders including, for example, connective tissue disease, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis, autoimmune pulmonary inflammation, Guillain-Barre syndrome, autoimmune thyroiditis, insulin dependent diabetes mellitis, myasthenia gravis, graft-versus-host disease and autoimmune inflammatory eye disease. Such a protein or polypeptide may also to be useful in the treatment of allergic reactions and conditions, such as asthma (particularly allergic asthma) or other respiratory problems. Other conditions, in which immune suppression is desired (including, for example, organ transplantation), may also be treatable using the protein or polypeptide.

[0302] Using the proteins or polypeptides of the invention it may also be possible to regulate immune responses either up or down. Down regulation may involve inhibiting or blocking an immune response already in progress or may involve preventing the induction of an immune response. The functions of activated T-cells may be inhibited by suppressing T cell responses or by inducing specific tolerance in T cells, or both. Immunosuppression of T cell responses is generally an active non-antigen-specific process which requires continuous exposure of the T cells to the suppressive agent. Tolerance, which involves inducing non-responsiveness or anergy in T cells, is distinguishable from immunosuppression in that it is generally antigen-specific and persists after the end of exposure to the tolerizing agent. Operationally, tolerance can be demonstrated by the lack of a T cell response upon reexposure to specific antigen in the absence of the tolerizing agent.

[0303] Down regulating or preventing one or more antigen functions (including without limitation B lymphocyte antigen functions, such as, for example, B7 costimulation), e.g., preventing high level lymphokine synthesis by activated T cells, will be useful in situations of tissue, skin and organ transplantation and in graft-versus-host disease (GVHD). For example, blockage of T cell function should result in reduced tissue destruction in tissue transplantation. Typically, in tissue transplants, rejection of the transplant is initiated through its recognition as foreign by T cells, followed by an immune reaction that destroys the transplant. The administration of a molecule which inhibits or blocks interaction of a B7 lymphocyte antigen with its natural ligand(s) on immune cells (such as a soluble, monomeric form of a peptide having B7-2 activity alone or in conjunction with a monomeric form of a peptide having an activity of another B lymphocyte antigen (e.g., B7-1, B7-3) or blocking antibody), prior to transplantation, can lead to the binding of the molecule to the natural ligand(s) on the immune cells without transmitting the corresponding costimulatory signal. Blocking B lymphocyte antigen function in this matter prevents cytokine synthesis by immune cells, such as T cells, and thus acts as an immunosuppressant. Moreover, the lack of costimulation may also be sufficient to anergize the T cells, thereby inducing tolerance in a subject. Induction of long-term tolerance by B lymphocyte antigen-blocking reagents may avoid the necessity of repeated administration of these blocking reagents. To achieve sufficient immunosuppression or tolerance in a subject, it may also be necessary to block the function of a combination of B lymphocyte antigens.

[0304] The efficacy of particular blocking reagents in preventing organ transplant rejection or GVHD can be assessed using animal models that are predictive of efficacy in humans. Examples of appropriate systems which can be used include allogeneic cardiac grafts in rats and xenogeneic pancreatic islet cell grafts in mice, both of which have been used to examine the immunosuppressive effects of CTLA4Ig fusion proteins *in vivo* as described in

Lenschow *et al., Science* **257**:789-792 (1992) and Turka *et al., Proc. Natl. Acad. Sci USA,* **89**:11102-11105 (1992). In addition, murine models of GVHD (see Paul ed., *Fundamental Immunology,* Raven Press, New York, 1989, pp. 846-847) can be used to determine the effect of blocking B lymphocyte antigen function *in vivo* on the development of that disease.

[0305] Blocking antigen function may also be therapeutically useful for treating autoimmune diseases. Many autoimmune disorders are the result of inappropriate activation of T cells that are reactive against self tissue and which promote the production of cytokines and autoantibodies involved in the pathology of the diseases. Preventing the activation of autoreactive T cells may reduce or eliminate disease symptoms. Administration of reagents which block costimulation of T cells by disrupting receptor/ligand interactions of B lymphocyte antigens can be used to inhibit T cell activation and prevent production of autoantibodies or T cell-derived cytokines which potentially involved in the disease process. Additionally, blocking reagents may induce antigen-specific tolerance of autoreactive T cells which could lead to long-term relief from the disease. The efficacy of blocking reagents in preventing or alleviating autoimmune disorders can be determined using a number of well-characterized animal models of human autoimmune diseases. Examples include murine experimental autoimmune encephalitis, systemic lupus erythmatosis in MRL/pr/pr mice or NZB hybrid mice, murine autoimmuno collagen arthritis, diabetes mellitus in OD mice and BB rats, and murine experimental myasthenia gravis (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 840-856).

[0306] Upregulation of an antigen function (preferably a B lymphocyte antigen function), as a means of up regulating immune responses, may also be useful in therapy. Upregulation of immune responses may involve either enhancing an existing immune response or eliciting an initial immune response as shown by the following examples. For instance, enhancing an immune response through stimulating B lymphocyte antigen function may be useful in cases of viral infection. In addition, systemic viral diseases such as influenza, the common cold, and encephalitis might be alleviated by the administration of stimulatory form of B lymphocyte antigens systemically.

[0307] Alternatively, antiviral immune responses may be enhanced in an infected patient by removing T cells from the patient, costimulating the T cells *in vitro* with viral antigen-pulsed APCs either expressing the proteins or polypeptides described above or together with a stimulatory form of the protein or polypeptide and reintroducing the *in vitro* primed T cells into the patient. The infected cells would now be capable of delivering a costimulatory signal to T cells *in vivo,* thereby activating the T cells.

[0308] In another application, upregulation or enhancement of antigen function (preferably B lymphocyte antigen function) may be useful in the induction of tumor immunity. Tumor cells (e.g., sarcoma, melanoma, lymphoma, leukemia, neuroblastoma, carcinoma) transfected with one of the above-described nucleic acids encoding a protein or polypeptide can be administered to a subject to overcome tumor-specific tolerance in the subject. If desired, the tumor cell can be transfected to express a combination of peptides. For example, tumor cells obtained from a patient can be transfected ex vivo with an expression vector directing the expression of a peptide having B7-2-like activity alone, or in conjunction with a peptide having B7-1-like activity and/or B7-3-like activity. The transfected tumor cells are returned to the patient to result in expression of the peptides on the surface of the transfected cell. Alternatively, gene therapy techniques can be used to target a tumor cell for transfection *in vivo.*

[0309] The presence of the protein or polypeptide encoded by the nucleic acids described above having the activity of a B lymphocyte antigen(s) on the surface of the tumor cell provides the necessary costimulation signal to T cells to induce a T cell mediated immune response against the transfected tumor cells. In addition, tumor cells which lack or which fail to reexpress sufficient amounts of MHC class I or MHC class II molecules can be transfected with nucleic acids encoding all or a portion of (e.g., a cytoplasmic-domain truncated portion) of an MHC class I a chain and β_2 microglobulin or an MHC class II α chain and an MHC class II β chain to thereby express MHC class I or MHC class II proteins on the cell surface, respectively. Expression of the appropriate MHC class I or class II molecules in conjunction with a peptide having the activity of a B lymphocyte antigen (e.g., B7-1, B7-2, B7-3) induces a T cell mediated immune response against the transfected tumor cell. Optionally, a nucleic acid encoding an antisense construct which blocks expression of an MHC class II associated protein, such as the invariant chain, can also be cotransfected with a DNA encoding a protein or polypeptide having the activity of a B lymphocyte antigen to promote presentation of tumor associated antigens and induce tumor specific immunity. Thus, the induction of a T cell mediated immune response in a human subject may be sufficient to overcome tumor-specific tolerance in the subject. Alternatively, as described in more detail below, nucleic acids encoding these immune system regulator proteins or polypeptides or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins as desired.

EXAMPLE 24

Assaying the Expressed Proteins or Polypeptides for Hematopoiesis Regulating Activity

[0310] The proteins or polypeptides encoded by the nucleic acids described above may also be evaluated for their hematopoiesis regulating activity. For example, the effect of the proteins or polypeptides on embryonic stem cell differentiation may be evaluated. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Johansson *et al. Cell. Biol.* **15**:141-151, 1995; Keller *et al., Mol. Cell. Biol.* **13**:473-486, 1993; McClanahan *et al., Blood* **81**:2903-2915, 1993.

[0311] The proteins or polypeptides encoded by the nucleic acids described above may also be evaluated for their influence on the lifetime of stem cells and stem cell differentiation. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Freshney, M.G.

Methylcellulose Colony Forming Assays, in Culture of Hematopoietic Cells. R.I. Freshney, et al. Eds. pp. 265-268, Wiley-Liss, Inc., New York, NY. 1994; Hirayama et al., *Proc. Natl. Acad. Sci. USA* 89:5907-5911, 1992; McNiece, I.K. and Briddell, R.A. Primitive Hematopoietic Colony Forming Cells with High Proliferative Potential, in Culture of Hematopoietic Cells. R.I. Freshney, et al. eds. Vol pp. 23-39, Wiley-Liss, Inc., New York, NY. 1994; Neben et al., *Experimental Hematology* 22:353-359, 1994; Ploemacher, R.E. Cobblestone Area Forming Cell Assay, In Culture of Hematopoietic Cells. R.I. Freshney, et al. Eds. pp. 1-21, Wiley-Liss, Inc., New York, NY. 1994; Spooncer, E., Dexter, M. and Allen, T. Long Term Bone Marrow Cultures in the Presence of Stromal Cells, in Culture of Hematopoietic Cells. R.I. Freshney, et al. Eds. pp. 163-179, Wiley-Liss, Inc., New York, NY. 1994; and Sutherland, H.J. Long Term Culture Initiating Cell Assay, in Culture of Hematopoietic Cells. R.I. Freshney, et al. Eds. pp. 139-162, Wiley-Liss, Inc., New York, NY. 1994.

[0312] Those proteins or polypeptides which exhibit hematopoiesis regulatory activity may then be formulated as pharmaceuticals and used to treat clinical conditions in which regulation of hematopoeisis is beneficial. For example, a protein or polypeptide of the present invention may be useful in regulation of hematopoiesis and, consequently, in the treatment of myeloid or lymphoid cell deficiencies. Even marginal biological activity in support of colony forming cells or of factor-dependent cell lines indicates involvement in regulating hematopoiesis, e.g. in supporting the growth and proliferation of erythroid progenitor cells alone or in combination with other cytokines, thereby indicating utility, for example, in treating various anemias or for use in conjunction with irradiation/chemotherapy to stimulate the production of erythroid precursors and/or erythroid cells; in supporting the growth and proliferation of myeloid cells such as granulocytes and monocytes/macrophages (i.e., traditional CSF activity) useful, for example, in conjunction with chemotherapy to prevent or treat consequent myelo-suppression; in supporting the growth and proliferation of megakaryocytes and consequently of platelets thereby allowing prevention or treatment of various platelet disorders such as thrombocytopenia, and generally for use in place of or complimentary to platelet transfusions; and/or in supporting the growth and proliferation of hematopoietic stem cells which are capable of maturing to any and all of the above-mentioned hematopoietic cells and therefore find therapeutic utility in various stem cell disorders (such as those usually treated with transplantion, including, without limitation, aplastic anemia and paroxysmal nocturnal hemoglobinuria), as well as in repopulating the stem cell compartment post irradiation/chemotherapy, either in-vivo or ex-vivo (i.e., in conjunction with bone marrow transplantation or with peripheral progenitor cell transplantation (homologous or heterologous)) as normal cells or genetically manipulated for gene therapy. Alternatively, as described in more detail below, nucleic acids encoding these proteins or polypeptides or nucleic acids regulating the expression of these proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins as desired.

EXAMPLE 25

Assaying the Expressed Proteins or Polypeptides for Regulation of Tissue Growth

[0313] The proteins or polypeptides encoded by the nucleic acids described above may also be evaluated for their effect on tissue growth. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in International Patent Publication No. WO95/16035, International Patent Publication No. WO95/05846 and International Patent Publication No. WO91/07491.

[0314] Assays for wound healing activity include, without limitation, those described in: Winter, *Epidermal Wound Healing*, pps. 71-112 (Maibach, H1 and Rovee, DT, eds.), Year Book Medical Publishers, Inc., Chicago, as modified by Eaglstein and Mertz, J. Invest. Dermatol 71:382-84 (1978).

[0315] Those proteins or polypeptides which are involved in the regulation of tissue growth may then be formulated as pharmaceuticals and used to treat clinical conditions in which regulation of tissue growth is beneficial. For example, a protein or polypeptide may have utility in compositions used for bone, cartilage, tendon, ligament and/or nerve tissue growth or regeneration, as well as for wound healing and tissue repair and replacement, and in the treatment of burns, incisions and ulcers.

[0316] A protein or polypeptide encoded by the nucleic acids described above which induces cartilage and/or bone growth in circumstances where bone is not normally formed, has application in the healing of bone fractures and cartilage damage or defects in humans and other animals. Such a preparation employing a protein or polypeptide of the invention may have prophylactic use in closed as well as open fracture reduction and also in the improved fixation of artificial joints. *De novo* bone synthesis induced by an osteogenic agent contributes to the repair of congenital, trauma induced, or oncologic resection induced craniofacial defects, and also is useful in cosmetic plastic surgery.

[0317] A protein or polypeptide of this invention may also be used in the treatment of periodontal disease, and in other tooth repair processes. Such agents may provide an environment to attract bone-forming cells, stimulate growth of bone-forming cells or induce differentiation of progenitors of bone-forming cells. A protein of the invention may also be useful in the treatment of osteoporosis or osteoarthritis, such as through stimulation of bone and/or cartilage repair or by blocking inflammation or processes of tissue destruction (collagenase activity, osteoclast activity, etc.) mediated by inflammatory processes.

[0318] Another category of tissue regeneration activity that may be attributable to the proteins or polypeptides encoded by the nucleic acids described above is tendon/ligament formation. A protein or polypeptide encoded by the nucleic acids described above, which induces tendon/ligament-like tissue or other tissue formation in circumstances where such tissue is not normally formed, has application in the healing of tendon or ligament tears, deformities and other tendon or ligament defects in humans and other animals. Such a preparation employing a tendon/ligament-

like tissue inducing protein may have prophylactic use in preventing damage to tendon or ligament tissue, as well as use in the improved fixation of tendon or ligament to bone or other tissues, and in repairing defects to tendon or ligament tissue. De novo tendon/ligament-like tissue formation induced by a protein or polypeptide of the present invention contributes to the repair of tendon or ligaments defects of congenital, traumatic or other origin and is also useful in cosmetic plastic surgery for attachment or repair of tendons or ligaments. The proteins or polypeptides of the present invention may provide an environment to attract tendon- or ligament-forming cells, stimulate growth of tendon- or ligament-forming cells, induce differentiation of progenitors of tendon- or ligament-forming cells, or induce growth of tendon/ligament cells or progenitors ex vivo for return *in vivo* to effect tissue repair. The proteins or polypeptides of the invention may also be useful in the treatment of tendinitis, carpal tunnel syndrome and other tendon or ligament defects. The therapeutic compositions may also include an appropriate matrix and/or sequestering agent as a carrier as is well known in the art.

[0319] The proteins or polypeptides of the present invention may also be useful for proliferation of neural cells and for regeneration of nerve and brain tissue, i.e., for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders, which involve degeneration, death or trauma to neural cells or nerve tissue. More specifically, a protein or polypeptide may be used in the treatment of diseases of the peripheral nervous system, such as peripheral nerve injuries, peripheral neuropathy and localized neuropathies, and central nervous system diseases, such as Alzheimer's, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and Shy-Drager syndrome. Further conditions which may be treated in accordance with the present invention include mechanical and traumatic disorders, such as spinal cord disorders, head trauma and cerebrovascular diseases such as stroke. Peripheral neuropathies resulting from chemotherapy or other medical therapies may also be treatable using a protein or polypeptide of the invention.

[0320] Proteins or polypeptides of the invention may also be useful to promote better or faster closure of non-healing wounds, including without limitation pressure ulcers, ulcers associated with vascular insufficiency, surgical and traumatic wounds, and the like.

[0321] It is expected that a protein or polypeptide of the present invention may also exhibit activity for generation or regeneration of other tissues, such as organs (including, for example, pancreas, liver, intestine, kidney, skin, endothelium) muscle (smooth, skeletal or cardiac) and vascular (including vascular endothelium) tissue, or for promoting the growth of cells comprising such tissues. Part of the desired effects may be by inhibition or modulation of fibrotic scarring to allow normal tissue to generate. A protein or polypeptide of the invention may also exhibit angiogenic activity.

[0322] A protein or polypeptide of the present invention may also be useful for gut protection or regeneration and treatment of lung or liver fibrosis, reperfusion injury in various tissues, and conditions resulting from systemic cytokine damage.

[0323] A protein or polypeptide of the present invention may also be useful for promoting or inhibiting differentiation of tissues described above from precursor tissues or cells; or for inhibiting the growth of tissues described above.

[0324] Alternatively, as described in more detail below, nucleic acids encoding tissue growth regulating activity proteins or polypeptides or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins as desired.

EXAMPLE 26

Assaying the Expressed Proteins or Polypeptides for Regulation of Reproductive Hormones

[0325] The proteins or polypeptides of the present invention may also be evaluated for their ability to regulate reproductive hormones, such as follicle stimulating hormone. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Vale *et al., Endocrinol.***91**:562-572, 1972; Ling *et al., Nature* **321**:779-782, 1986; Vale *et al., Nature***321**:776-779, 1986; Mason *et al., Nature* **318**:659-663, 1985; Forage *et al., Proc. Natl. Acad. Sci. USA* **83**:3091-3095, 1986. Chapter 6.12 in *Current Protocols in Immunology,* J.E. Coligan *et al.* Eds. Greene Publishing Associates and Wiley-Interscience ; Taub *et al. J. Clin. Invest.* **95**:1370-1376, 1995; Lind *et al. APMIS***103**:140-146, 1995; Muller *et al. Eur. J. Immunol.* **25**:1744-1748; Gruber *et al. J. Immunol.* **152**:5860-5867, 1994; Johnston *et al., J Immunol.* **153**:1762-1768, 1994.

[0326] Those proteins or polypeptides which exhibit activity as reproductive hormones or regulators of cell movement may then be formulated as pharmaceuticals and used to treat clinical conditions in which regulation of reproductive hormones are beneficial. For example, a protein or polypeptide may exhibit activin- or inhibin-related activities. Inhibins are characterized by their ability to inhibit the release of follicle stimulating hormone (FSH), while activins are characterized by their ability to stimulate the release of FSH. Thus, a protein or polypeptide of the present invention, alone or in heterodimers with a member of the inhibin a family, may be useful as a contraceptive based on the ability of inhibins to decrease fertility in female mammals and decrease spermatogenesis in male mammals. Administration of sufficient amounts of other inhibins can induce infertility in these mammals. Alternatively, the protein or polypeptide of the invention, as a homodimer or as a heterodimer with other protein subunits of the inhibin-B group, may be useful as a fertility inducing therapeutic, based upon the ability of activin molecules in stimulating FSH release from cells of the anterior pituitary. See, for example, United States Patent 4,798,885. A protein or polypeptide of the invention may also be useful for advancement of the onset of fertility in sexually immature mammals, so as to increase the lifetime reproductive performance of domestic

animals such as cows, sheep and pigs.

[0327] Alternatively, as described in more detail below, nucleic acids encoding reproductive hormone regulating activity proteins or polypeptides or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 27

Assaying the Expressed Proteins or Polypeptides For Chemotactic/Chemokinetic Activity

[0328] The proteins or polypeptides of the present invention may also be evaluated for chemotactic/chemokinetic activity. For example, a protein or polypeptide of the present invention may have chemotactic or chemokinetic activity (e.g., act as a chemokine) for mammalian cells, including, for example, monocytes, fibroblasts, neutrophils, T-cells, mast cells, eosinophils, epithelial and/or endothelial cells. Chemotactic and chemokinetic proteins or polypeptides can be used to mobilize or attract a desired cell population to a desired site of action. Chemotactic or chemokinetic proteins or polypeptides provide particular advantages in treatment of wounds and other trauma to tissues, as well as in treatment of localized infections. For example, attraction of lymphocytes, monocytes or neutrophils to tumors or sites of infection may result in improved immune responses against the tumor or infecting agent.

[0329] A protein or polypeptide has chemotactic activity for a particular cell population if it can stimulate, directly or indirectly, the directed orientation or movement of such cell population. Preferably, the protein or polypeptide has the ability to directly stimulate directed movement of cells. Whether a particular protein or polypeptide has chemotactic activity for a population of cells can be readily determined by employing such protein or polypeptide in any known assay for cell chemotaxis.

[0330] The activity of a protein or polypeptide of the invention may, among other means, be measured by the following methods:

[0331] Assays for chemotactic activity (which will identify proteins or polypeptides that induce or prevent chemotaxis) consist of assays that measure the ability of a protein or polypeptide to induce the migration of cells across a membrane as well as the ability of a protein or polypeptide to induce the adhesion of one cell population to another cell population. Suitable assays for movement and adhesion include, without limitation, those described in: *Current Protocols in Immunology*, Ed by J.E. Coligan, A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience,Chapter **6.12**: 6.12.1-6.12.28; Taub *et al. J. Clin. Invest.* **95**:1370-1376, 1995; Lind *et al. APMIS* **103**:140-146, 1995; Mueller *et al., Eur. J. Immunol.* **25**:1744-1748; Gruber *et al. J. Immunol.* **152**:5860-5867, 1994; Johnston *et al. J. Immunol.*, **153**:1762-1768, 1994.

EXAMPLE 28

Assaying the Expressed Proteins or Polypeptides for Regulation of Blood Clotting

[0332] The proteins or polypeptides of the present invention may also be evaluated for their effects on blood clotting. Numerous assays for such activity are familiar to those skilled in the art, including the assays disclosed in the following references: Linet *et al., J. Clin. Pharmacol.* **26**:131-140, 1986; Burdick *et al., Thrombosis Res.* **45**:413-419, 1987; Humphrey *et al., Fibrinolysis* **5**:71-79 (1991); Schaub,*Prostaglandins* **35**:467-474, 1988.

[0333] Those proteins or polypeptides which are involved in the regulation of blood clotting may then be formulated as pharmaceuticals and used to treat clinical conditions in which regulation of blood clotting is beneficial. For example, a protein or polypeptide of the invention may also exhibit hemostatic or thrombolytic activity. As a result, such a protein or polypeptide is expected to be useful in treatment of various coagulations disorders (including hereditary disorders, such as hemophilias) or to enhance coagulation and other hemostatic events in treating wounds resulting from trauma, surgery or other causes. A protein or polypeptide of the invention may also be useful for dissolving or inhibiting formation of thromboses and for treatment and prevention of conditions resulting therefrom (such as infarction of cardiac and central nervous system vessels (e.g., stroke)). Alternatively, as described in more detail below, nucleic acids encoding blood clotting activity proteins or polypeptides or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 29

Assaying the Expressed Proteins or Polypeptides for Involvement in Receptor/Ligand Interactions

[0334] The proteins or polypeptides of the present invention may also be evaluated for their involvement in receptor/ligand interactions. Numerous assays for such involvement are familiar to those skilled in the art, including the assays disclosed in the following references: Chapter 7. 7.28.1-7.28.22) in *Current Protocols in Immunology*, J.E. Coligan *et al.* Eds. Greene Publishing Associates and Wiley-Interscience; Takai *et al., Proc. Natl. Acad. Sci. USA* **84**:6864-6868, 1987; Bierer *et al., J. Exp. Med.* **168**:1145-1156, 1988; Rosenstein *et al., J. Exp. Med.* **169**:149-160, 1989; Stoltenborg *et al., J. Immunol. Methods* **175**:59-68, 1994, Stitt *et al., Cell* **80**:661-670, 1995; Gyuris *et al., Cell* **75**:791-803, 1993.

[0335] For example, the proteins or polypeptides of the present invention may also demonstrate activity as receptors, receptor ligands or inhibitors or agonists of receptor/ligand interactions. Examples of such receptors and ligands include, without limitation, cytokine receptors and their ligands, receptor kinases and their ligands, receptor phosphatases and their ligands, receptors involved in cell-cell interactions and their ligands (including without limitation, cellular adhesion molecules (such as selectins, integrins and their ligands) and receptor/ligand pairs involved in antigen presentation, antigen recognition and development of cellular and humoral immune responses) . Receptors and ligands are also useful for screening of potential peptide or small molecule inhibitors of the relevant receptor/ligand interaction. A protein or polypeptide of the present invention (including, without limitation, fragments of receptors and ligands) maybe useful as inhibitors of receptor/ligand interactions. Alternatively, as described in more detail below, nucleic acids encoding proteins or polypeptides involved in receptor/ligand interactions or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 30

Assaying the Proteins or Polypeptides for Anti-Inflammatory Activity

[0336] The proteins or polypeptides of the present invention may also be evaluated for anti-inflammatory activity. The anti-inflammatory activity may be achieved by providing a stimulus to cells involved in the inflammatory response, by inhibiting or promoting cell-cell interactions (such as, for example, cell adhesion), by inhibiting or promoting chemotaxis of cells involved in the inflammatory process, inhibiting or promoting cell extravasation, or by stimulating or suppressing production of other factors which more directly inhibit or promote an inflammatory response. Proteins or polypeptides exhibiting such activities can be used to treat inflammatory conditions including chronic or acute conditions, including without limitation inflammation associated with infection (such as septic shock, sepsis or systemic inflammatory response syndrome), ischemiareperfusioninury, endotoxin lethality, arthritis, complement-mediated hyperacute rejection, nephritis, cytokine- or chemokine-induced lung injury, inflammatory bowel disease, Crohn's disease or resulting from over production of cytokines such as TNF or IL-1. Proteins or polypeptides of the invention may also be useful to treat anaphylaxis and hypersensitivity to an antigenic substance or material. Alternatively, as described in more detail below, nucleic acids encoding anti-inflammatory activity proteins or polypeptides or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 31

Assaying the Expressed Proteins or Polypeptides for Tumor Inhibition Activity

[0337] The proteins or polypeptides of the present invention may also be evaluated for tumor inhibition activity. In addition to the activities described above for immunological treatment or prevention of tumors, a protein or polypeptide of the invention may exhibit other anti-tumor activities. A protein or polypeptide may inhibit tumor growth directly or indirectly (such as, for example, via ADCC). A protein or polypeptide may exhibit its tumor inhibitory activity by acting on tumor tissue or tumor precursor tissue, by inhibiting formation of tissues necessary to support tumor growth (such as, for example, by inhibiting angiogenesis), by causing production of other factors, agents or cell types which inhibit tumor growth, or by suppressing, eliminating or inhibiting factors, agents or cell types which promote tumor growth. . Alternatively, as described in more detail below, nucleic acids encoding proteins or polypeptides with tumor inhibition activity or nucleic acids regulating the expression of such proteins or polypeptides may be introduced into appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

[0338] A protein or polypeptide of the invention may also exhibit one or more of the following additional activities or effects: inhibiting the growth, infection or function of, or killing, infectious agents, including, without limitation, bacteria, viruses, fungi and other parasites; effecting (suppressing or enhancing) bodily characteristics, including, without limitation, height, weight, hair color, eye color, skin, fat to lean ratio or other tissue pigmentation, or organ or body part size or shape (such as, for example, breast augmentation or diminution, change in bone form or shape); effecting biorhythms or circadian cycles or rhythms; effecting the fertility of male or female subjects; effecting the metabolism, catabolism, anabolism, processing, utilization, storage or elimination of dietary fat, lipid, protein, carbohydrate, vitamins, minerals, cofactors or other nutritional factors or component(s); effecting behavioral characteristics, including, without limitation, appetite, libido, stress, cognition (including cognitive disorders), depression (including depressive disorders) and violent behaviors; providing analgesic effects or other pain reducing effects; promoting differentiation and growth of embryonic stem cells in lineages other than hematopoietic lineages; hormonal or endocrine activity; in the case of enzymes, correcting deficiencies of the enzyme and treating deficiency-related diseases; treatment of hyperproliferative disorders (such as, for example, psoriasis); immunoglobulin-like activity (such as, for example, the ability to bind antigens or complement); and the ability to act as an antigen in a vaccine composition to raise an immune response against such protein or another material or entity which is cross-reactive with such protein. Alternatively, as described in more detail below, nucleic acids encoding proteins or polypeptides involved in any of the above mentioned activities or nucleic acids regulating the expression of such proteins may be introduced into

appropriate host cells to increase or decrease the expression of the proteins or polypeptides as desired.

EXAMPLE 32

Identification of Proteins or Polypeptides which Interact with Proteins or Polypeptides of the Present Invention

[0339] Proteins or polypeptides which interact with the proteins or polypeptides of the present invention, such as receptor proteins, may be identified using two hybrid systems such as the Matchmaker Two Hybrid System 2 (Catalog No. K1604-1, Clontech). As described in the manual accompanying the kit, nucleic acids encoding the proteins or polypeptides of the present invention, are inserted into an expression vector such that they are in frame with DNA encoding the DNA binding domain of the yeast transcriptional activator GAL4. cDNAs in a cDNA library which encode proteins or polypeptides which might interact with the proteins or polypeptides of the present invention are inserted into a second expression vector such that they are in frame with DNA encoding the activation domain of GAL4. The two expression plasmids are transformed into yeast and the yeast are plated on selection medium which selects for expression of selectable markers on each of the expression vectors as well as GALA dependent expression of the HIS3 gene. Transformants capable of growing on medium lacking histidine are screened for GAL4 dependent lacZ expression. Those cells which are positive in both the histidine selection and the lacZ assay contain plasmids encoding proteins or polypeptides which interact with the proteins or polypeptides of the present invention.

[0340] Alternatively, the system described in Lustig *et al., Methods in Enzymology* **283**: 83-99 (1997), may be used for identifying molecules which interact with the proteins or polypeptides of the present invention. In such systems, *in vitro* transcription reactions are performed on a pool of vectors containing nucleic acid inserts which encode the proteins or polypeptides of the present invention. The nucleic acid inserts are cloned downstream of a promoter which drives *in vitro* transcription. The resulting pools of mRNAs are introduced into *Xenopus laevis* oocytes. The oocytes are then assayed for a desired activity.

[0341] Alternatively, the pooled *in vitro* transcription products produced as described above may be translated *in vitro*. The pooled *in vitro* translation products can be assayed for a desired activity or for interaction with a known protein or polypeptide.

[0342] Proteins, polypeptides or other molecules interacting with proteins or polypeptides of the present invention can be found by a variety of additional techniques. In one method, affinity columns containing the protein or polypeptide of the present invention can be constructed. In some versions, of this method the affinity column contains chimeric proteins in which the protein or polypeptide of the present invention is fused to glutathione S-transferase. A mixture of cellular proteins or pool of expressed proteins as described above and is applied to the affinity column. Molecules interacting with the protein or polypeptide attached to the column can then be isolated and analyzed on 2-D electrophoresis gel as described in Ramunsen *et al. Electrophoresis,* **18**, 588-598 (1997). Alternatively, the molecules retained on the affinity column can be purified by electrophoresis based methods and sequenced. The same method can be used to isolate antibodies, to screen phage display products, or to screen phage display human antibodies.

[0343] Molecules interacting with the proteins or polypeptides of the present invention can also be screened by using an Optical Biosensor as described in Edwards & Leatherbarrow, *Analytical Biochemistry,* **246**, 1-6 (1997). The main advantage of the method is that it allows the determination of the association rate between the protein or polypeptide and other interacting molecules. Thus, it is possible to specifically select interacting molecules with a high or low association rate. Typically a target molecule is linked to the sensor surface (through a carboxymethl dextran matrix) and a sample of test molecules is placed in contact with the target molecules. The binding of a test molecule to the target molecule causes a change in the refractive index and/ or thickness. This change is detected by the Biosensor provided it occurs in the evanescent field (which extend a few hundred nanometers from the sensor surface). In these screening assays, the target molecule can be one of the proteins or polypeptides of the present invention and the test sample can be a collection of proteins, polypeptides or other molecules extracted from tissues or cells, a pool of expressed proteins, combinatorial peptide and/ or chemical libraries, or phage displayed peptides. The tissues or cells from which the test molecules are extracted can originate from any species.

[0344] In other methods, a target protein or polypeptide is immobilized and the test population is a collection of unique proteins or polypeptides of the present invention.

[0345] To study the interaction of the proteins or polypeptides of the present invention with drugs, the microdialysis coupled to HPLC method described by Wang *et al., Chromatographia,* **44**, 205-208(1997) or the affinity capillary electrophoresis method described by Busch *et al., J. Chromatogr.* **777**:311-328 (1997) can be used.

[0346] The system described in U.S. Patent No. 5,654,150 may also be used to identify molecules which interact with the proteins or polypeptides of the present invention. In this system, pools of nucleic acids encoding the proteins or polypeptides of the present invention are transcribed and translated *in vitro* and the reaction products are assayed for interaction with a known polypeptide or antibody.

[0347] It will be appreciated by those skilled in the art that the proteins or polypeptides of the present invention may be assayed for numerous activities in addition to those specifically enumerated above. For example, the expressed proteins or polypeptides may be evaluated for applications involving control and regulation of inflammation, tumor proliferation or metastasis, infection, or other clinical conditions. In addition, the proteins or polypeptides may be useful as nutritional agents or cosmetic agents.

[0348] The proteins or polypeptides of the present invention may be used to generate antibodies capable of specifically binding to the proteins or polypeptides of the present invention. The antibodies may be monoclonal

antibodies or polyclonal antibodies. As used herein, "antibody" refers to a polypeptide or group of polypeptides which are comprised of at least one binding domain, where a binding domain is formed from the folding of variable domains of an antibody molecule to form three-dimensional binding spaces with an internal surface shape and charge distribution complementary to the features of an antigenic determinant of an antigen., which allows an immunological reaction with the antigen. Antibodies include recombinant proteins comprising the binding domains, as wells as fragments, including Fab, Fab', $F(ab)_2$, and $F(ab')_2$ fragments.

[0349] As used herein, an "antigenic determinant" is the portion of an antigen molecule, that determines the specificity of the antigen-antibody reaction. An "epitope" refers to an antigenic determinant of a polypeptide. An epitope can comprise as few as 3 amino acids in a spatial conformation which is unique to the epitope. Generally an epitope consists of at least 6 such amino acids, and more usually at least 8-10 such amino acids. Methods for determining the amino acids which make up an epitope include x-ray crystallography, 2-dimensional nuclear magnetic resonance, and epitope mapping e.g. the Pepscan method described by H. Mario Geysen et al. 1984. Proc. Natl. Acad. Sci. U.S.A. 81:3998-4002; PCT Publication No. WO 84/03564; and PCT Publication No. WO 84/03506.

[0350] In some embodiments, the antibodies may be capable of specifically binding to a protein or polypeptide encoded by EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. In some embodiments, the antibody may be capable of binding an antigenic determinant or an epitope in a protein or polypeptide encoded by EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids.

[0351] In other embodiments, the antibodies may be capable of specifically binding to an EST-related polypeptide, fragment of an EST-related polypeptide, positional segment of an EST-related polypeptide or fragment of a positional segment of an EST-related polypeptide. In some embodiments, the antibody may be capable of binding an antigenic determinant or an epitope in an EST-related polypeptide, fragment of an EST-related polypeptide, positional segment of an EST-related polypeptide or fragment of a positional segment of an EST-related polypeptide.

[0352] In the case of secreted proteins, the antibodies may be capable of binding a full-length protein encoded by a nucleic acid of the present invention, a mature protein (i.e. the protein generated by cleavage of the signal peptide) encoded by a nucleic acid of the present invention, or a signal peptide encoded by a nucleic acid of the present invention.

EXAMPLE 33

Production of an Antibody to a Human Polypeptide or Protein

[0353] The above described EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or nucleic acids encoding EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides are operably linked to promoters and introduced into cells as described above.

[0354] In the case of secreted proteins, nucleic acids encoding the full protein (i.e. the mature protein and the signal peptide), nucleic acids encoding the mature protein (i.e. the protein generated by cleavage of the signal peptide), or nucleic acids encoding the signal peptide are operably linked to promoters and introduced into cells as described above.

[0355] The encoded proteins or polypeptides are then substantially purified or isolated as described above. The concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few µg/ml. Monoclonal or polyclonal antibody to the protein or polypeptide can then be prepared as follows:

1. Monoclonal Antibody Production by Hybridoma Fusion

[0356] Monoclonal antibody to epitopes of any of the proteins or polypeptides identified and isolated as described can be prepared from murine hybridomas according to the classical method of Kohler, and Milstein, *Nature* **256**:495 (1975) or derivative methods thereof. Briefly, a mouse is repetitively inoculated with a few micrograms of the selected protein or peptides derived therefrom over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as Elisa, as originally described by Engvall, *Meth. Enzymol.* **70**:419 (1980). Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. *et al.* in *Basic Methods in Molecular Biology* Elsevier, New York. Section 21-2.

2. Polyclonal Antibody Production by Immunization

[0357] Polyclonal antiserum containing antibodies to heterogenous epitopes of a single protein or polypeptide

can be prepared by immunizing suitable animals with the expressed protein or peptides derived therefrom, which can be unmodified or modified to enhance immunogenicity. Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. For example, small molecules tend to be less immunogenic than others and may require the use of carriers and adjuvant. Also, host animals response vary depending on site of inoculations and doses, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. An effective immunization protocol for rabbits can be found in Vaitukaitis. *et al.J. Clin. Endocrinol. Metab.***33**:988-991 (1971).

[0358] Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall. See, for example, Ouchterlony, *et al.,* Chap. 19 in: *Handbook of Experimental Immunology* D. Wier (ed) Blackwell (1973). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μM). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: *Manual of Clinical Immunology.* 2d Ed. (Rose and Friedman, Eds.) Amer. Soc. For Microbiol., Washington, D.C. (1980).

[0359] Antibody preparations prepared according to either of the above protocols are useful in a variety of contexts. In particular, the antibodies may be used in immunoaffinity chromatography techniques such as those described below to facilitate large scale isolation, purification, or enrichment of the proteins or polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or for the isolation, purification or enrichment of EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides.

[0360] In the case of secreted proteins, the antibodies may be used for the isolation, purification, or enrichment of the full protein (i.e. the mature protein and the signal peptide), the mature protein (i.e. the protein generated by cleavage of the signal peptide), or the signal peptide are operably linked to promoters and introduced into cells as described above.

[0361] Additionally, the antibodies may be used in immunoaffinity chromatography techniques such as those described below to isolate, purify, or enrich polypeptides which have been linked to the proteins or polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or to isolate, purify, or enrich EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides.

[0362] The antibodies may also be used to determine the cellular localization of polypeptides encoded by the proteins or polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or the cellular localization of EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides.

[0363] In addition, the antibodies may also be used to determine the cellular localization of polypeptides which have been linked to the proteins or polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or polypeptides which have been linked EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides .

[0364] The antibodies may also be used in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they may also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample or to identify the type of tissue present in a biological sample. The antibodies may also be used in therapeutic compositions for killing cells expressing the protein or reducing the levels of the protein in the body.

V. Use of 5'ESTs and Consensus Contigated 5' ESTs or Sequences Obtainable Therefrom or Portions Thereof as Reagents

[0365] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used as reagents in isolation procedures, diagnostic assays, and forensic procedures. For example, sequences from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids, may be detectably labeled and used as probes to isolate other sequences capable of hybridizing to them. In addition, the he EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used to design PCR primers to be used in isolation, diagnostic, or forensic procedures.

1. Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids in isolation, diagnostic and forensic procedures

EXAMPLE 34

Preparation of PCR Primers and Amplification of DNA

[0366] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used to prepare PCR primers for a variety of applications, including isolation procedures for cloning nucleic acids capable of hybridizing to such sequences, diagnostic techniques and forensic techniques. In some embodiments, the PCR primers at least 10, 15, 18, 20, 23, 25, 28, 30, 40, or 50 nucleotides in length. In some embodiments, the PCR primers may be more than 30 bases in length. It is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. A variety of PCR techniques are familiar to those skilled in the art. For a review of PCR technology, see Molecular Cloning to Genetic Engineering White, B.A. Ed. *in Methods in Molecular Biology* **67:** Humana Press, Totowa 1997. In each of these PCR procedures, PCR primers on either side of the nucleic acid sequences to be amplified are added to a suitably prepared nucleic acid sample along with dNTPs and a thermostable polymerase such as Taq polymerase, Pfu polymerase, or Vent polymerase. The nucleic acid in the sample is denatured and the PCR primers are specifically hybridized to complementary nucleic acid sequences in the sample. The hybridized primers are extended. Thereafter, another cycle of denaturation, hybridization, and extension is initiated. The cycles are repeated multiple times to produce an amplified fragment containing the nucleic acid sequence between the primer sites.

EXAMPLE 35

Use of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids as probes

[0367] Probes derived from EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be labeled with detectable labels familiar to those skilled in the art, including radioisotopes and non-radioactive labels, to provide a detectable probe. The detectable probe may be single stranded or double stranded and may be made using techniques known in the art, including *in vitro* transcription, nick translation, or kinase reactions. A nucleic acid sample containing a sequence capable of hybridizing to the labeled probe is contacted with the labeled probe. If the nucleic acid in the sample is double stranded, it may be denatured prior to contacting the probe. In some applications, the nucleic acid sample may be immobilized on a surface such as a nitrocellulose or nylon membrane. The nucleic acid sample may comprise nucleic acids obtained from a variety of sources, including genomic DNA, cDNA libraries, RNA, or tissue samples.

[0368] Procedures used to detect the presence of nucleic acids capable of hybridizing to the detectable probe include well known techniques such as Southern blotting, Northern blotting, dot blotting, colony hybridization, and plaque hybridization. In some applications, the nucleic acid capable of hybridizing to the labeled probe may be cloned into vectors such as expression vectors, sequencing vectors, or *in vitro* transcription vectors to facilitate the characterization and expression of the hybridizing nucleic acids in the sample. For example, such techniques may be used to isolate and clone sequences in a genomic library or cDNA library which are capable of hybridizing to the detectable probe as described in Example 18 above.

[0369] PCR primers made as described in Example 34 above may be used in forensic analyses, such as the DNA fingerprinting techniques described in Examples 36-40 below. Such analyses may utilize detectable probes or primers based on the sequences of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids.

EXAMPLE 36

Forensic Matching by DNA Sequencing

[0370] In one exemplary method, DNA samples are isolated from forensic specimens of, for example, hair, semen, blood or skin cells by conventional methods. A panel of PCR primers based on a number of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is then utilized in accordance with Example 34 to amplify DNA of approximately 100-200 bases in length from the forensic specimen. Corresponding sequences are obtained from a test subject. Each of these identification DNAs is then sequenced using standard techniques, and a simple database comparison determines the differences, if any, between the sequences from the subject and those from the sample. Statistically significant differences between the suspect's DNA sequences and those from the sample conclusively prove a lack of identity. This lack of identity can be proven, for example, with only one sequence. Identity, on the other hand, should be demonstrated with a large number of sequences, all matching. Preferably, a minimum of 50 statistically identical sequences of 100 bases in length are used to prove identity between the suspect and the sample.

EXAMPLE 37

Positive Identification by DNA Sequencing

[0371] The technique outlined in the previous example may also be used on a larger scale to provide a unique fingerprint-type identification of any individual. In this technique, primers are prepared from a large number of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. Preferably, 20 to 50 different primers are used. These primers are used to obtain a

corresponding number of PCR-generated DNA segments from the individual in question in accordance with Example 34. Each of these DNA segments is sequenced, using the methods set forth in Example 36. The database of sequences generated through this procedure uniquely identifies the individual from whom the sequences were obtained. The same panel of primers may then be used at any later time to absolutely correlate tissue or other biological specimen with that individual.

EXAMPLE 38

Southern Blot Forensic Identification

[0372] The procedure of Example 37 is repeated to obtain a panel of at least 10 amplified sequences from an individual and a specimen. Preferably, the panel contains at least 50 amplified sequences. More preferably, the panel contains 100 amplified sequences. In some embodiments, the panel contains 200 amplified sequences. This PCR-generated DNA is then digested with one or a combination of, preferably, four base specific restriction enzymes. Such enzymes are commercially available and known to those of skill in the art. After digestion, the resultant gene fragments are size separated in multiple duplicate wells on an agarose gel and transferred to nitrocellulose using Southern blotting techniques well known to those with skill in the art. For a review of Southern blotting see Davis *et al.* (Basic Methods in Molecular Biology, 1986, Elsevier Press. pp 62-65).

[0373] A panel of probes based on the sequences of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are radioactively or colorimetrically labeled using methods known in the art, such as nick translation or end labeling, and hybridized to the Southern blot using techniques known in the art (Davis *et al.,* supra). Preferably, the probe is at least 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 nucleotides in length. Preferably, the probes are at least 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 nucleotides in length. In some embodiments, the probes are oligonucleotides which are 40 nucleotides in length or less.

[0374] Preferably, at least 5 to 10 of these labeled probes are used, and more preferably at least about 20 or 30 are used to provide a unique pattern. The resultant bands appearing from the hybridization of a large sample of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids will be a unique identifier. Since the restriction enzyme cleavage will be different for every individual, the band pattern on the Southern blot will also be unique. Increasing the number of probes will provide a statistically higher level of confidence in the identification since there will be an increased number of sets of bands used for identification.

EXAMPLE 39

Dot Blot Identification Procedure

[0375] Another technique for identifying individuals using the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids disclosed herein utilizes a dot blot hybridization technique.

[0376] Genomic DNA is isolated from nuclei of subject to be identified. Probes are prepared that correspond to at least 10, preferably 50 sequences from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. The probes are used to hybridize to the genomic DNA through conditions known to those in the art. The oligonucleotides are end labeled with P^{32}à using polynucleotide kinase (Pharmacia). Dot Blots are created by spotting the genomic DNA onto nitrocellulose or the like using a vacuum dot blot manifold (BioRad, Richmond California). The nitrocellulose filter containing the genomic sequences is baked or UV linked to the filter, prehybridized and hybridized with labeled probe using techniques known in the art (Davis et *al., supra*). The ^{32}àP labeled DNA fragments are sequentially hybridized with successively stringent conditions to detect minimal differences between the 30 bp sequence and the DNA. Tetramethylammonium chloride is useful for identifying clones containing small numbers of nucleotide mismatches (Wood *et al., Proc. Natl. Acad. Sci. USA* **82(6)**:1585-1588 (1985)). A unique pattern of dots distinguishes one individual from another individual.

[0377] EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids can be used as probes in the following alternative fingerprinting technique. In some embodiments, the probes are oligonucleotides which are 40 nucleotides in length or less.

[0378] Preferably, a plurality of probes having sequences from different EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are used in the alternative fingerprinting technique. Example 40 below provides a representative alternative fingerprinting procedure in which the probes are derived from EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids.

EXAMPLE 40

Alternative "Fingerprint" Identification Technique

[0379] Oligonucleotides are prepared from a large number, e.g. 50, 100, or 200, EST-related nucleic acids,

positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids using commercially available oligonucleotide services such as Genset, Paris, France. Preferably, the oligonucleotides are at least 10, 15, 18, 20, 23, 25 28, or 30 nucleotides in length. However, in some embodiments, the oligonucleotides may be more than 30 nucleotides in length.

[0380] Cell samples from the test subject are processed for DNA using techniques well known to those with skill in the art. The nucleic acid is digested with restriction enzymes such as EcoRI and Xbal. Following digestion, samples are applied to wells for electrophoresis. The procedure, as known in the art, may be modified to accommodate polyacrylamide electrophoresis, however in this example, samples containing 5 ug of DNA are loaded into wells and separated on 0.8% agarose gels. The gels are transferred onto nitrocellulose using standard Southern blotting techniques.

[0381] 10 ng of each of the oligonucleotides are pooled and end-labeled with P^{32}à. The nitrocellulose is prehybridized with blocking solution and hybridized with the labeled probes. Following hybridization and washing, the nitrocellulose filter is exposed to X-Omat AR X-ray film. The resulting hybridization pattern will be unique for each individual.

[0382] It is additionally contemplated within this example that the number of probe sequences used can be varied for additional accuracy or clarity.

[0383] In addition to their applications in forensics and identification, EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be mapped to their chromosomal locations. Example 41 below describes radiation hybrid (RH) mapping of human chromosomal regions using EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. Example 42 below describes a representative procedure for mapping EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to their locations on human chromosomes. Example 43 below describes mapping of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids on metaphase chromosomes by Fluorescence In Situ Hybridization (FISH).

2. Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids in Chromosome Mapping

EXAMPLE 41

Radiation hybrid mapping of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to the human genome

[0384] Radiation hybrid (RH) mapping is a somatic cell genetic approach that can be used for high resolution mapping of the human genome. In this approach, cell lines containing one or more human chromosomes are lethally irradiated, breaking each chromosome into fragments whose size depends on the radiation dose. These fragments are rescued by fusion with cultured rodent cells, yielding subclones containing different portions of the human genome. This technique is described by Benham *et al.* (*Genomics* **4**:509-517, 1989) and Cox *et al.*, (*Science* **250**:245-250, 1990) . The random and independent nature of the subclones permits efficient mapping of any human genome marker. Human DNA isolated from a panel of 80-100 cell lines provides a mapping reagent for ordering EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. In this approach, the frequency of breakage between markers is used to measure distance, allowing construction of fine resolution maps as has been done using conventional ESTs (Schuler *et al., Science* **274**:540-546, 1996).

[0385] RH mapping has been used to generate a high-resolution whole genome radiation hybrid map of human chromosome 17q22-q25.3 across the genes for growth hormone (GH) and thymidine kinase (TK) (Foster *et al., Genomics* **33**:185-192, 1996), the region surrounding the Gorlin syndrome gene (Obermayr *et al., Eur. J. Hum. Genet.* **4**:242-245, 1996), 60 loci covering the entire short arm of chromosome 12 (Raeymaekers *et al., Genomics* **29**:170-178, 1995), the region of human chromosome 22 containing the neurofibromatosis type 2 locus (Frazer *et al., Genomics* **14**:574-584, 1992) and 13 loci on the long arm of chromosome 5 (Warrington *et al., Genomics* **11**:701-708, 1991).

EXAMPLE 42

Mapping of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to Human Chromosomes using PCR techniques

[0386] EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be assigned to human chromosomes using PCR based methodologies. In such approaches, oligonucleotide primer pairs are designed from EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to minimize the chance of amplifying through an intron. Preferably, the oligonucleotide primers are 18-23 bp in length and are designed for PCR amplification. The creation of PCR primers from known sequences is well known to those with skill in the art. For a review of PCR technology see Erlich. in PCR Technology; Principles and Applications for DNA Amplification.

1992. W.H. Freeman and Co., New York.

[0387] The primers are used in polymerase chain reactions (PCR) to amplify templates from total human genomic DNA. PCR conditions are as follows: 60 ng of genomic DNA is used as a template for PCR with 80 ng of each oligonucleotide primer, 0.6 unit of Taq polymerase, and 1 µCu of a 32P-labeled deoxycytidine triphosphate. The PCR is performed in a microplate thermocycler (Techne) under the following conditions: 30 cycles of 94°C, 1.4 min; 55°C, 2 min; and 72°C, 2 min; with a final extension at 72°C for 10 min. The amplified products are analyzed on a 6% polyacrylamide sequencing gel and visualized by autoradiography. If the length of the resulting PCR product is identical to the distance between the ends of the primer sequences in the 5'EST from which the primers are derived, then the PCR reaction is repeated with DNA templates from two panels of human-rodent somatic cell hybrids, BIOS PCRable DNA (BIOS Corporation) and NIGMS Human-Rodent Somatic Cell Hybrid Mapping Panel Number 1 (NIGMS, Camden, NJ).

[0388] PCR is used to screen a series of somatic cell hybrid cell lines containing defined sets of human chromosomes for the presence of a given 5'EST. DNA is isolated from the somatic hybrids and used as starting templates for PCR reactions using the primer pairs from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. Only those somatic cell hybrids with chromosomes containing the human gene corresponding to the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids will yield an amplified fragment. The 5'ESTs are assigned to a chromosome by analysis of the segregation pattern of PCR products from the somatic hybrid DNA templates. The single human chromosome present in all cell hybrids that give rise to an amplified fragment is the chromosome containing that EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. For a review of techniques and analysis of results from somatic cell gene mapping experiments. (See Ledbetter et al., Genomics 6:475-481 (1990)).

[0389] Alternatively, the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be mapped to individual chromosomes using FISH as described in Example 43 below.

EXAMPLE 43

Mapping of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to Chromosomes Using

Fluorescence *In Situ* Hybridization

[0390] Fluorescence in situ hybridization allows the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to be mapped to a particular location on a given chromosome. The chromosomes to be used for fluorescence in situ hybridization techniques may be obtained from a variety of sources including cell cultures, tissues, or whole blood.

[0391] In a preferred embodiment, chromosomal localization of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are obtained by FISH as described by Cherif *et al.* (*Proc. Natl. Acad. Sci. U.S.A.*, **87**:6639-6643, 1990). Metaphase chromosomes are prepared from phytohemagglutinin (PHA)-stimulated blood cell donors. PHA-stimulated lymphocytes from healthy males are cultured for 72 h in RPMI-1640 medium. For synchronization, methotrexate (10 µM) is added for 17 h, followed by addition of 5-bromodeoxyuridine (5-BrdU, 0.1 mM) for 6 h. Colcemid (1 µg/ml) is added for the last 15 min before harvesting the cells. Cells are collected, washed in RPMI, incubated with a hypotonic solution of KCl (75 mM) at 37°C for 15 min and fixed in three changes of methanol:acetic acid (3:1). The cell suspension is dropped onto a glass slide and air dried. The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is labeled with biotin-16 dUTP by nick translation according to the manufacturer's instructions (Bethesda Research Laboratories, Bethesda, MD), purified using a Sephadex G-50 column (Pharmacia, Upsala, Sweden) and precipitated. Just prior to hybridization, the DNA pellet is dissolved in hybridization buffer (50% formamide, 2 X SSC, 10% dextran sulfate, 1 mg/ml sonicated salmon sperm DNA, pH 7) and the probe is denatured at 70°C for 5-10 min.

[0392] Slides kept at -20°C are treated for 1 h at 37°C with RNase A (100 µg/ml), rinsed three times in 2 X SSC and dehydrated in an ethanol series. Chromosome preparations are denatured in 70% formamide, 2 X SSC for 2 min at 70°C, then dehydrated at 4°C. The slides are treated with proteinase K (10 µg/100 ml in 20 mM Tris-HCl, 2 mM $CaCl_2$) at 37°C for 8 min and dehydrated. The hybridization mixture containing the probe is placed on the slide, covered with a coverslip, sealed with rubber cement and incubated overnight in a humid chamber at 37°C. After hybridization and post-hybridization washes, the biotinylated probe is detected by avidin-FITC and amplified with additional layers of biotinylated goat anti-avidin and avidin-FITC. For chromosomal localization, fluorescent R-bands are obtained as previously described (Cherif *et al., supra.*). The slides are observed under a LEICA fluorescence microscope (DMRXA). Chromosomes are counterstained with propidium iodide and the fluorescent signal of the probe appears as two symmetrical yellow-green spots on both chromatids of the fluorescent R-band chromosome (red). Thus, a particular EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be localized to a particular cytogenetic R-band on a given chromosome. Once the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids have been assigned to particular chromosomes using the techniques described in Examples 41-43 above, they may be utilized to construct a high resolution map of the chromosomes on which they

are located or to identify the chromosomes in a sample.

EXAMPLE 44

Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to Construct or Expand Chromosome Maps

[0393] Chromosome mapping involves assigning a given unique sequence to a particular chromosome as described above. Once the unique sequence has been mapped to a given chromosome, it is ordered relative to other unique sequences located on the same chromosome. One approach to chromosome mapping utilizes a series of yeast artificial chromosomes (YACs) bearing several thousand long inserts derived from the chromosomes of the organism from which the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are obtained. This approach is described in Ramaiah Nagaraja *et al., Genome Research* 7:210-222, March 1997. Briefly, in this approach each chromosome is broken into overlapping pieces which are inserted into the YAC vector. The YAC inserts are screened using PCR or other methods to determine whether they include the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids whose position is to be determined. Once an insert has been found which includes the 5'EST, the insert can be analyzed by PCR or other methods to determine whether the insert also contains other sequences known to be on the chromosome or in the region from which the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids was derived. This process can be repeated for each insert in the YAC library to determine the location of each of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids relative to one another and to other known chromosomal markers. In this way, a high resolution map of the distribution of numerous unique markers along each of the organisms chromosomes may be obtained.

[0394] As described in Example 45 below EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may also be used to identify genes associated with a particular phenotype, such as hereditary disease or drug response.

3. Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids Gene Identification

EXAMPLE 45

Identification of genes associated with hereditary diseases or drug response

[0395] This example illustrates an approach useful for the association of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids with particular phenotypic characteristics. In this example, a particular EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is used as a test probe to associate that EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids with a particular phenotypic characteristic.

[0396] EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are mapped to a particular location on a human chromosome using techniques such as those described in Examples 41 and 42 or other techniques known in the art. A search of Mendelian Inheritance in Man (V. McKusick, *Mendelian Inheritance in Man* (available on line through Johns Hopkins University Welch Medical Library) reveals the region of the human chromosome which contains the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to be a very gene rich region containing several known genes and several diseases or phenotypes for which genes have not been identified. The gene corresponding to this EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids thus becomes an immediate candidate for each of these genetic diseases.

[0397] Cells from patients with these diseases or phenotypes are isolated and expanded in culture. PCR primers from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are used to screen genomic DNA, mRNA or cDNA obtained from the patients. EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids that are not amplified in the patients can be positively associated with a particular disease by further analysis. Alternatively, the PCR analysis may yield fragments of different lengths when the samples are derived from an individual having the phenotype associated with the disease than when the sample is derived from a healthy individual, indicating that the gene containing the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be responsible for the genetic disease.

VI. Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to Construct Vectors

[0398] The present EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may also be used to construct secretion vectors capable of directing the secretion of the proteins encoded by genes therein. Such secretion vectors may facilitate the purification or enrichment of the proteins encoded by genes inserted therein by reducing the number of background proteins from which the desired protein must be purified or enriched. Exemplary secretion vectors are described in Example 46 below.

1. Construction of secretion vectors

EXAMPLE 46

Construction of Secretion Vectors

[0399] The secretion vectors of the present invention include a promoter capable of directing gene expression in the host cell, tissue, or organism of interest. Such promoters include the Rous Sarcoma Virus promoter, the SV40 promoter, the human cytomegalovirus promoter, and other promoters familiar to those skilled in the art.

[0400] A signal sequence from one of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is operably linked to the promoter such that the mRNA transcribed from the promoter will direct the translation of the signal peptide. Preferably, the signal sequence is from one of the nucleic acids of SEQ ID NOs.:24-4100. The host cell, tissue, or organism may be any cell, tissue, or organism which recognizes the signal peptide encoded by the signal sequence in the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. Suitable hosts include mammalian cells, tissues or organisms, avian cells, tissues, or organisms, insect cells, tissues or organisms, or yeast.

[0401] In addition, the secretion vector contains cloning sites for inserting genes encoding the proteins which are to be secreted. The cloning sites facilitate the cloning of the insert gene in frame with the signal sequence such that a fusion protein in which the signal peptide is fused to the protein encoded by the inserted gene is expressed from the mRNA transcribed from the promoter. The signal peptide directs the extracellular secretion of the fusion protein.

[0402] The secretion vector may be DNA or RNA and may integrate into the chromosome of the host, be stably maintained as an extrachromosomal replicon in the host, be an artificial chromosome, or be transiently present in the host. Preferably, the secretion vector is maintained in multiple copies in each host cell. As used herein, multiple copies means at least 2, 5, 10, 20, 25, 50 or more than 50 copies per cell. In some embodiments, the multiple copies are maintained extrachromosomally. In other embodiments, the multiple copies result from amplification of a chromosomal sequence.

[0403] Many nucleic acid backbones suitable for use as secretion vectors are known to those skilled in the art, including retroviral vectors, SV40 vectors, Bovine Papilloma Virus vectors, yeast integrating plasmids, yeast episomal plasmids, yeast artificial chromosomes, human artificial chromosomes, P element vectors, baculovirus vectors, or bacterial plasmids capable of being transiently introduced into the host.

[0404] The secretion vector may also contain a polyA signal such that the polyA signal is located downstream of the gene inserted into the secretion vector.

[0405] After the gene encoding the protein for which secretion is desired is inserted into the secretion vector, the secretion vector is introduced into the host cell, tissue, or organism using calcium phosphate precipitation, DEAE-Dextran, electroporation, liposome-mediated transfection, viral particles or as naked DNA. The protein encoded by the inserted gene is then purified or enriched from the supernatant using conventional techniques such as ammonium sulfate precipitation, immunoprecipitation, immunoaffinitychromatography, size exclusion chromatography, ion exchange chromatography, and HPLC. Alternatively, the secreted protein may be in a sufficiently enriched or pure state in the supernatant or growth media of the host to permit it to be used for its intended purpose without further enrichment.

[0406] The signal sequences may also be inserted into vectors designed for gene therapy. In such vectors, the signal sequence is operably linked to a promoter such that mRNA transcribed from the promoter encodes the signal peptide. A cloning site is located downstream of the signal sequence such that a gene encoding a protein whose secretion is desired may readily be inserted into the vector and fused to the signal sequence. The vector is introduced into an appropriate host cell. The protein expressed from the promoter is secreted extracellularly, thereby producing a therapeutic effect.

EXAMPLE 47

Fusion Vectors

[0407] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used to construct fusion vectors for the expression of chimeric polypeptides. The chimeric polypeptides comprise a first polypeptide portion and a second polypeptide portion. In the fusion vectors of the present invention, nucleic acids encoding the first polypeptide portion and the second polypeptide portion are joined in frame with one another so as to generate a nucleic acid encoding the

chimeric polypeptide. The nucleic acid encoding the chimeric polypeptide is operably linked to a promoter which directs the expression of an mRNA encoding the chimeric polypeptide. The promoter may be in any of the expression vectors described herein including those described in Examples 20 and 46.

[0408] Preferably, the fusion vector is maintained in multiple copies in each host cell. In some embodiments, the multiple copies are maintained extrachromosomally. In other embodiments, the multiple copies result from amplification of a chromosomal sequence.

[0409] The first polypeptide portion may comprise any of the polypeptides encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. In some embodiments, the first polypeptide portion may be one of the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides.

[0410] The second polypeptide portion may comprise any polypeptide of interest. In some embodiments, the second polypeptide portion may comprise a polypeptide having a detectable enzymatic activity such as green fluorescent protein or β galactosidase. Chimeric polypeptides in which the second polypeptide portion comprises a detectable polypeptide may be used to determine the intracellular localization of the first polypeptide portion. In such procedures, the fusion vector encoding the chimeric polypeptide is introduced into a host cell under conditions which facilitate the expression of the chimeric polypeptide. Where appropriate, the cells are treated with a detection reagent which is visible under the microscope following a catalytic reaction with the detectable polypeptide and the cellular location of the detection reagent is determined. For example, if the polypeptide having a detectable enzymatic activity is β galactosidase, the cells may be treated with Xgal. Alternatively, where the detectable polypeptide is directly detectable without the addition of a detection reagent, the intracellular location of the chimeric polypeptide is determined by performing microscopy under conditions in which the dectable polypeptide is visible. For example, if the detectable polypeptide is green fluorescent protein or a modified version thereof, microscopy is performed by exposing the host cells to light having an appropriate wavelength to cause the green fluorescent protein or modified version thereof to fluoresce.

[0411] Alternatively, the second polypeptide portion may comprise a polypeptide whose isolation, purification, or enrichment is desired. In such embodiments, the isolation, purification, or enrichment of the second polypeptide portion may be achieved by performing the immunoaffinity chromatography procedures described below using an immunoaffinity column having an antibody directed against the first polypeptide portion coupled thereto.

[0412] The proteins encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides may also be used to generate antibodies as explained in Examples 20 and 33 in order to identify the tissue type or cell species from which a sample is derived as described in Example 48.

EXAMPLE 48

Identification of Tissue Types or Cell Species by Means of Labeled Tissue Specific Antibodies

[0413] Identification of specific tissues is accomplished by the visualization of tissue specific antigens by means of antibody preparations according to Examples 20 and 33 which are conjugated, directly or indirectly to a detectable marker. Selected labeled antibody species bind to their specific antigen binding partner in tissue sections, cell suspensions, or in extracts of soluble proteins from a tissue sample to provide a pattern for qualitative or semi-qualitative interpretation.

[0414] Antisera for these procedures must have a potency exceeding that of the native preparation, and for that reason, antibodies are concentrated to a mg/ml level by isolation of the gamma globulin fraction, for example, by ion-exchange chromatography or by ammonium sulfate fractionation. Also, to provide the most specific antisera, unwanted antibodies, for example to common proteins, must be removed from the gamma globulin fraction, for example by means of insoluble immunoabsorbents, before the antibodies are labeled with the marker. Either monoclonal or heterologous antisera is suitable for either procedure.

1. Immunohistochemical Techniques

[0415] Purified, high-titer antibodies, prepared as described above, are conjugated to a detectable marker, as described, for example, by Fudenberg, H., Chap. 26 in: *Basic 503 Clinical Immunology,* 3rd Ed. Lange, Los Altos, California (1980) or Rose,. *et al.,* Chap. **12** in: *Methods in Immunodiagnosis,* 2d Ed. John Wiley and Sons, New York (1980).

[0416] A fluorescent marker, either fluorescein or rhodamine, is preferred, but antibodies can also be labeled with an enzyme that supports a color producing reaction with a substrate, such as horseradish peroxidase. Markers can be added to tissue-bound antibody in a second step, as described below. Alternatively, the specific antitissue antibodies can be labeled with ferritin or other electron dense particles, and localization of the ferritin coupled antigen-antibody complexes achieved by means of an electron microscope. In yet another approach, the antibodies are radiolabeled, with, for example ^{125}Al, and detected by overlaying the antibody treated preparation with photographic emulsion.

[0417] Preparations to carry out the procedures can comprise monoclonal or polyclonal antibodies to a single

protein or peptide identified as specific to a tissue type, for example, brain tissue, or antibody preparations to several antigenically distinct tissue specific antigens can be used in panels, independently or in mixtures, as required.

[0418] Tissue sections and cell suspensions are prepared for immunohistochemical examination according to common histological techniques. Multiple cryostat sections (about 4 µm, unfixed) of the unknown tissue and known control, are mounted and each slide covered with different dilutions of the antibody preparation. Sections of known and unknown tissues should also be treated with preparations to provide a positive control, a negative control, for example, pre-immune sera, and a control for non-specific staining, for example, buffer.

[0419] Treated sections are incubated in a humid chamber for 30 min at room temperature, rinsed, then washed in buffer for 30-45 min. Excess fluid is blotted away, and the marker developed.

[0420] If the tissue specific antibody was not labeled in the first incubation, it can be labeled at this time in a second antibody-antibody reaction, for example, by adding fluorescein- or enzyme-conjugated antibody against the immunoglobulin class of the antiserum-producing species, for example, fluorescein labeled antibody to mouse IgG. Such labeled sera are commercially available.

[0421] The antigen found in the tissues by the above procedure can be quantified by measuring the intensity of color or fluorescence on the tissue section, and calibrating that signal using appropriate standards.

2. Identification of Tissue Specific Soluble Proteins

[0422] The visualization of tissue specific proteins and identification of unknown tissues from that procedure is carried out using the labeled antibody reagents and detection strategy as described for immunohistochemistry; however the sample is prepared according to an electrophoretic technique to distribute the proteins extracted from the tissue in an orderly array on the basis of molecular weight for detection.

[0423] A tissue sample is homogenized using a Virtis apparatus; cell suspensions are disrupted by Dounce homogenization or osmotic lysis, using detergents in either case as required to disrupt cell membranes, as is the practice in the art. Insoluble cell components such as nuclei, microsomes, and membrane fragments are removed by ultracentrifugation, and the soluble protein-containing fraction concentrated if necessary and reserved for analysis.

[0424] A sample of the soluble protein solution is resolved into individual protein species by conventional SDS polyacrylamide electrophoresis as described, for example, by Davis,L. *et al.,* Section **19-2** in: *Basic Methods in Molecular Biology* (P. Leder, ed), Elsevier, New York (1986), using a range of amounts of polyacrylamide in a set of gels to resolve the entire molecular weight range of proteins to be detected in the sample. A size marker is run in parallel for purposes of estimating molecular weights of the constituent proteins. Sample size for analysis is a convenient volume of from 5 to 55 µl, and containing from about 1 to 100 µg protein. An aliquot of each of the resolved proteins is transferred by blotting to a nitrocellulose filter paper, a process that maintains the pattern of resolution. Multiple copies are prepared. The procedure, known as Western Blot Analysis, is well described in Davis, L. *et al., supra* Section 19-3. One set of nitrocellulose blots is stained with Coomassie Blue dye to visualize the entire set of proteins for comparison with the antibody bound proteins. The remaining nitrocellulose filters are then incubated with a solution of one or more specific antisera to tissue specific proteins prepared as described in Examples 20 and 33. In this procedure, as in procedure A above, appropriate positive and negative sample and reagent controls are run.

[0425] In either procedure described above a detectable label can be attached to the primary tissue antigen-primary antibody complex according to various strategies and permutations thereof. In a straightforward approach, the primary specific antibody can be labeled; alternatively, the unlabeled complex can be bound by a labeled secondary anti-IgG antibody. In other approaches, either the primary or secondary antibody is conjugated to a biotin molecule, which can, in a subsequent step, bind an avidin conjugated marker. According to yet another strategy, enzyme labeled or radioactive protein A, which has the property of binding to any IgG, is bound in a final step to either the primary or secondary antibody.

EXAMPLE 49

Immunohistochemical Localization of Polypeptides

[0426] The antibodies prepared as described in Examples 20 and 33 above may be utilized to determine the cellular location of a polypeptide. The polypeptide may be any of the polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or the polypeptide may be one of the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides. In some embodiments, the polypeptide may be a chimeric polypeptide such as those encoded by the fusion vectors of Example 47.

[0427] Cells expressing the polypeptide to be localized are applied to a microscope slide and fixed using any of the procedures typically employed in immunohistochemical localization techniques, including the methods described in *Current Protocols in Molecular Biology,* John Wiley and Sons, Inc. 1997. Following a washing step, the cells are contacted with the antibody. In some embodiments, the antibody is conjugated to a detectable marker as described above to facilitate detection. Alternatively, in some embodiments, after the cells have been contacted with an antibody to the polypeptide to be localized, a secondary antibody which has been conjugated to a detectable marker is placed in contact with the antibody against the polypeptide to be localized.

[0428] Thereafter, microscopy is performed under conditions suitable for visualizing the cellular location of the polypeptide.

[0429] The visualization of tissue specific antigen binding at levels above those seen in control tissues to one or more tissue specific antibodies, directed against the polypeptides encoded by EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or antibodies against the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides, can identify tissues of unknown origin, for example, forensic samples, or differentiated tumor tissue that has metastasized to foreign bodily sites.

[0430] The antibodies of Example 20 and 33 may also be used in the immunoaffinity chromatography techniques described below to isolate, purify or enrich the polypeptides encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or to isolate, purify or enrich EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides. The immunoaffinity chromatography techniques described below may also be used to isolate, purify or enrich polypeptides which have been linked to the polypeptides encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or to isolate, purify or enrich polypeptides which have been linked to EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides.

EXAMPLE 50

Immunoaffinity Chromatography

[0431] Antibodies prepared as described above are coupled to a support. Preferably, the antibodies are monoclonal antibodies, but polyclonal antibodies may also be used. The support may be any of those typically employed in immunoaffinity chromatography, including Sepharose CL-4B (Pharmacia, Piscataway, NJ), Sepharose CL-2B (Pharmacia, Piscataway, NJ), Affi-gel 10 (Biorad, Richmond, CA), or glass beads.

[0432] The antibodies may be coupled to the support using any of the coupling reagents typically used in immunoaffinity chromatography, including cyanogen bromide. After coupling the antibody to the support, the support is contacted with a sample which contains a target polypeptide whose isolation, purification or enrichment is desired. The target polypeptide may be a polypeptide encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or the target polypeptide may be one of the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides. The target polypeptides may also be polypeptides which have been linked to the polypeptides encoded by the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or the target polypeptides may be polypeptides which have been linked to EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of positional segments of EST-related polypeptides using the fusion vectors described above.

[0433] Preferably, the sample is placed in contact with the support for a sufficient amount of time and under appropriate conditions to allow at least 50% of the target polypeptide to specifically bind to the antibody coupled to the support.

[0434] Thereafter, the support is washed with an appropriate wash solution to remove polypeptides which have non-specifically adhered to the support. The wash solution may be any of those typically employed in immunoaffinity chromatography, including PBS, Tris-lithium chloride buffer (0.1 M lysine base and 0.5M lithium chloride, pH 8.0), Tris-hydrochloride buffer (0.05M Tris-hydrochloride, pH 8.0), or Tris/Triton/NaCl buffer (50mM Tris.cl, pH 8.0 or 9.0, 0.1% Triton X-100, and 0.5MNaCl).

[0435] After washing, the specifically bound target polypeptide is eluted from the support using the high pH or low pH elution solutions typically employed in immunoaffinity chromatography. In particular, the elution solutions may contain an eluant such as triethanolamine, diethylamine, calcium chloride, sodium thiocyanate, potasssium bromide, acetic acid. or glycine. In some embodiments, the elution solution may also contain a detergent such as Triton X-100 or octyl-β-D-glucoside.

[0436] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may also be used to clone sequences located upstream of the 5'ESTs which are capable of regulating gene expression, including promoter sequences, enhancer sequences, and other upstream sequences which influence transcription or translation levels. Once identified and cloned, these upstream regulatory sequences may be used in expression vectors designed to direct the expression of an inserted gene in a desired spatial, temporal, developmental, or quantitative fashion. Example 51 describes a method for cloning sequences upstream of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids.

2. Identification of upstream sequences with promoting or regulatory activities

EXAMPLE 51

Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to Clone Upstream Sequences from Genomic DNA

[0437] Sequences derived from EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used to isolate the promoters of the corresponding genes using chromosome walking techniques. In one chromosome walking technique, which utilizes the GenomeWalker™ kit available from Clontech, five complete genomic DNA samples are each digested with a different restriction enzyme which has a 6 base recognition site and leaves a blunt end. Following digestion, oligonucleotide adapters are ligated to each end of the resulting genomic DNA fragments.

[0438] For each of the five genomic DNA libraries, a first PCR reaction is performed according to the manufacturer's instructions using an outer adapter primer provided in the kit and an outer gene specific primer. The gene specific primer should be selected to be specific for 5' EST of interest and should have a melting temperature, length, and location in the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids which is consistent with its use in PCR reactions. Each first PCR reaction contains 5ng of genomic DNA, 5 µl of 10X Tth reaction buffer, 0.2 mM of each dNTP, 0.2 µM each of outer adapter primer and outer gene specific primer, 1.1 mM of $Mg(OAc)_2$, and 1 µl of the Tth polymerase 50X mix in a total volume of 50 µl. The reaction cycle for the first PCR reaction is as follows: 1 min at 94°C / 2 sec at 94°C, 3 min at 72°C (7 cycles) / 2 sec at 94°C, 3 min at 67°C (32 cycles) / 5 min at 67°C.

[0439] The product of the first PCR reaction is diluted and used as a template for a second PCR reaction according to the manufacturer's instructions using a pair of nested primers which are located internally on the amplicon resulting from the first PCR reaction. For example, 5 µl of the reaction product of the first PCR reaction mixture may be diluted 180 times. Reactions are made in a 50 µl volume having a composition identical to that of the first PCR reaction except the nested primers are used. The first nested primer is specific for the adapter, and is provided with the GenomeWalker™ kit. The second nested primer is specific for the particular EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids for which the promoter is to be cloned and should have a melting temperature, length, and location in the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids which is consistent with its use in PCR reactions. The reaction parameters of the second PCR reaction are as follows: 1 min at 94°C / 2 sec at 94°C, 3 min at 72°C (6 cycles) / 2 sec at 94°C; 3 min at 67°C (25 cycles) / 5 min at - 67°C. The product of the second PCR reaction is purified, cloned, and sequenced using standard techniques.

[0440] Alternatively, two or more human genomic DNA libraries can be constructed by using two or more restriction enzymes. The digested genomic DNA is cloned into vectors which can be converted into single stranded, circular, or linear DNA. A biotinylated oligonucleotide comprising at least 15 nucleotides from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids sequence is hybridized to the single stranded DNA. Hybrids between the biotinylated oligonucleotide and the single stranded DNA containing the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are isolated as described above. Thereafter, the single stranded DNA containing the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is released from the beads and converted into double stranded DNA using a primer specific for the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids or a primer corresponding to a sequence included in the cloning vector. The resulting double stranded DNA is transformed into bacteria. cDNAs containing the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are identified by colony PCR or colony hybridization.

[0441] Once the upstream genomic sequences have been cloned and sequenced as described above, prospective promoters and transcription start sites within the upstream sequences may be identified by comparing the sequences upstream of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids with databases containing known transcription start sites, transcription factor binding sites, or promoter sequences.

[0442] In addition, promoters in the upstream sequences may be identified using promoter reporter vectors as described in Example 53.

EXAMPLE 53

Identification of Promoters in Cloned Upstream Sequences

[0443] The genomic sequences upstream of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are cloned into a suitable promoter reporter vector, such as the pSEAP-Basic, pSEAP-Enhancer, pβgal-Basic, pβgal-Enhancer, or pEGFP-1 Promoter Reporter vectors available from Clontech. Briefly, each of these promoter reporter vectors include multiple cloning sites positioned upstream of a reporter gene encoding a readily assayable protein such as secreted alkaline phosphatase, β galactosidase, or green fluorescent protein. The sequences upstream of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are inserted into the cloning sites upstream of the reporter gene in both orientations and introduced into an

appropriate host cell. The level of reporter protein is assayed and compared to the level obtained from a vector which lacks an insert in the cloning site. The presence of an elevated expression level in the vector containing the insert with respect to the control vector indicates the presence of a promoter in the insert. If necessary, the upstream sequences can be cloned into vectors which contain an enhancer for augmenting transcription levels from weak promoter sequences. A significant level of expression above that observed with the vector lacking an insert indicates that a promoter sequence is present in the inserted upstream sequence.

[0444] Appropriate host cells for the promoter reporter vectors may be chosen based on the results of the above described determination of expression patterns of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. For example, if the expression pattern analysis indicates that the mRNA corresponding to a particular EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids is expressed in fibroblasts, the promoter reporter vector may be introduced into a human fibroblast cell line.

[0445] Promoter sequences within the upstream genomic DNA may be further defined by constructing nested deletions in the upstream DNA using conventional techniques such as Exonuclease III digestion. The resulting deletion fragments can be inserted into the promoter reporter vector to determine whether the deletion has reduced or obliterated promoter activity. In this way, the boundaries of the promoters may be defined. If desired, potential individual regulatory sites within the promoter may be identified using site directed mutagenesis or linker scanning to obliterate potential transcription factor binding sites within the promoter individually or in combination. The effects of these mutations on transcription levels may be determined by inserting the mutations into the cloning sites in the promoter reporter vectors.

EXAMPLE 54

Cloning and Identification of Promoters

[0446] Using the method described in Example 51 above with 5' ESTs, sequences upstream of several genes were obtained. Using the primer pairs GGG AAG ATG GAG ATA GTA TTG CCT G (SEQ ID NO:15) and CTG CCA TGT ACA TGA TAG AGA GAT TC (SEQ ID NO: 16), the promoter having the internal designation P13H2 (SEQ ID NO:17) was obtained.

[0447] Using the primer pairs GTA CCA GGGG ACT GTG ACC ATT GC (SEQ ID NO:18) and CTG TGA CCA TTG CTC CCA AGA GAG (SEQ ID NO:19), the promoter having the internal designation P15B4 (SEQ ID NO:20) was obtained.

[0448] Using the primer pairs CTG GGA TGG AAG GCA CGG TA (SEQ ID NO:21) and GAG ACC ACA CAG CTA GAC AA (SEQ ID NO:22), the promoter having the internal designation P29B6 (SEQ ID NO:23) was obtained.

[0449] Figure 4 provides a schematic description of the promoters isolated and the way they are assembled with the corresponding 5' tags. The upstream sequences were screened for the presence of motifs resembling transcription factor binding sites or known transcription start sites using the computer program MatInspector release 2.0, August 1996.

[0450] Figure 5 describes the transcription factor binding sites present in each of these promoters. The columns labeled matrice provides the name of the MatInspector matrix used. The column labeled position provides the 5' position of the promoter site. Numeration of the sequence starts from the transcription site as determined by matching the genomic sequence with the 5' EST sequence. The column labeled "orientation" indicates the DNA strand on which the site is found, with the + strand being the coding strand as determined by matching the genomic sequence with the sequence of the 5' EST. The column labeled "score" provides the MatInspector score found for this site. The column labeled "length" provides the length of the site in nucleotides. The column labeled "sequence" provides the sequence of the site found.

[0451] Bacterial clones containing plasmids containing the promoter sequences described above described above are presently stored in the inventor's laboratories under the internal identification numbers provided above. The inserts may be recovered from the deposited materials by growing an aliquot of the appropriate bacterial clone in the appropriate medium. The plasmid DNA can then be isolated using plasmid isolation procedures familiar to those skilled in the art such as alkaline lysis minipreps or large scale alkaline lysis plasmid isolation procedures. If desired the plasmid DNA may be further enriched by centrifugation on a cesium chloride gradient, size exclusion chromatography, or anion exchange chromatography. The plasmid DNA obtained using these procedures may then be manipulated using standard cloning techniques familiar to those skilled in the art. Alternatively, a PCR can be done with primers designed at both ends of the inserted EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. The PCR product which corresponds to the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids can then be manipulated using standard cloning techniques familiar to those skilled in the art.

[0452] The promoters and other regulatory sequences located upstream of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used to design expression vectors capable of directing the expression of an inserted gene in a desired spatial, temporal, developmental, or quantitative manner. A promoter capable of directing the desired spatial, temporal, developmental, and quantitative patterns may be selected using the results of the expression analysis described above. For example, if a promoter which confers a high level of expression in muscle is desired, the

promoter sequence upstream of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids derived from an mRNA which are expressed at a high level in muscle, as determined by the methods above, may be used in the expression vector.

[0453] Preferably, the desired promoter is placed near multiple restriction sites to facilitate the cloning of the desired insert downstream of the promoter, such that the promoter is able to drive expression of the inserted gene. The promoter may be inserted in conventional nucleic acid backbones designed for extrachromosomal replication, integration into the host chromosomes or transient expression. Suitable backbones for the present expression vectors include retroviral backbones, backbones from eukaryotic episomes such as SV40 or Bovine Papilloma Virus, backbones from bacterial episomes, or artificial chromosomes.

[0454] Preferably, the expression vectors also include a polyA signal downstream of the multiple restriction sites for directing the polyadenylation of mRNA transcribed from the gene inserted into the expression vector.

[0455] Following the identification of promoter sequences using the procedures of Examples 51-54, proteins which interact with the promoter may be identified as described in Example 55 below.

EXAMPLE 55

Identification of Proteins Which Interact with Promoter Sequences, Upstream Regulatory Sequences, or mRNA

[0456] Sequences within the promoter region which are likely to bind transcription factors may be identified by homology to known transcription factor binding sites or through conventional mutagenesis or deletion analyses of reporter plasmids containing the promoter sequence. For example, deletions may be made in a reporter plasmid containing the promoter sequence of interest operably linked to an assayable reporter gene. The reporter plasmids carrying various deletions within the promoter region are transfected into an appropriate host cell and the effects of the deletions on expression levels is assessed. Transcription factor binding sites within the regions in which deletions reduce expression levels may be further localized using site directed mutagenesis, linker scanning analysis, or other techniques familiar to those skilled in the art.

[0457] Nucleic acids encoding proteins which interact with sequences in the promoter may be identified using one-hybrid systems such as those described in the manual accompanying the Matchmaker One-Hybrid System kit available from Clontech (Catalog No. K1603-1). Briefly, the Matchmaker One-hybrid system is used as follows. The target sequence for which it is desired to identify binding proteins is cloned upstream of a selectable reporter gene and integrated into the yeast genome. Preferably, multiple copies of the target sequences are inserted into the reporter plasmid in tandem. A library comprised of fusions between cDNAs to be evaluated for the ability to bind to the promoter and the activation domain of a yeast transcription factor, such as GAL4, is transformed into the yeast strain containing the integrated reporter sequence. The yeast are plated on selective media to select cells expressing the selectable marker linked to the promoter sequence. The colonies which grow on the selective media contain genes encoding proteins which bind the target sequence. The inserts in the genes encoding the fusion proteins are further characterized by sequencing. In addition, the inserts may be inserted into expression vectors or *in vitro* transcription vectors. Binding of the polypeptides encoded by the inserts to the promoter DNA may be confirmed by techniques familiar to those skilled in the art, such as gel shift analysis or DNAse protection analysis.

VII. Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids in Gene Therapy

[0458] The present invention also comprises the use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids in gene therapy strategies, including antisense and triple helix strategies as described in Examples 56 and 57 below. In antisense approaches, nucleic acid sequences complementary to an mRNA are hybridized to the mRNA intracellularly, thereby blocking the expression of the protein encoded by the mRNA. The antisense sequences may prevent gene expression through a variety of mechanisms. For example, the antisense sequences may inhibit the ability of ribosomes to translate the mRNA. Alternatively, the antisense sequences may block transport of the mRNA from the nucleus to the cytoplasm, thereby limiting the amount of mRNA available for translation. Another mechanism through which antisense sequences may inhibit gene expression is by interfering with mRNA splicing. In yet another strategy, the antisense nucleic acid may be incorporated in a ribozyme capable of specifically cleaving the target mRNA.

EXAMPLE 56

Preparation and Use of Antisense Oligonucleotides

[0459] The antisense nucleic acid molecules to be used in gene therapy may be either DNA or RNA sequences. They may comprise a sequence complementary to the sequence of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids. The antisense nucleic acids should have a length and melting temperature sufficient to permit formation of an intracellular duplex with sufficient stability to inhibit the expression of the mRNA in the duplex. Strategies for designing antisense nucleic acids suitable for use in gene therapy are disclosed in Green *et al.*, *Ann. Rev. Biochem.* **55**:569-597 (1986) and Izant and Weintraub, *Cell* **36**:1007-1015 (1984).

[0460] In some strategies, antisense molecules are obtained from a nucleotide sequence encoding a protein by

reversing the orientation of the coding region with respect to a promoter so as to transcribe the opposite strand from that which is normally transcribed in the cell. The antisense molecules may be transcribed using *in vitro* transcription systems such as those which employ T7 or SP6 polymerase to generate the transcript. Another approach involves transcription of the antisense nucleic acids *in vivo* by operably linking DNA containing the antisense sequence to a promoter in an expression vector.

[0461] Alternatively, oligonucleotides which are complementary to the strand normally transcribed in the cell may be synthesized *in vitro*. Thus, the antisense nucleic acids are complementary to the corresponding mRNA and are capable of hybridizing to the mRNA to create a duplex. In some embodiments, the antisense sequences may contain modified sugar phosphate backbones to increase stability and make them less sensitive to RNase activity. Examples of modifications suitable for use in antisense strategies are described by Rossi *et al., Pharmacol. Ther.* **50(2)**:245-254, (1991).

[0462] Various types of antisense oligonucleotides complementary to the sequence of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may be used. In one preferred embodiment, stable and semi-stable antisense oligonucleotides described in International Application No. PCT W094/23026 are used. In these molecules, the 3' end or both the 3' and 5' ends are engaged in intramolecular hydrogen bonding between complementary base pairs. These molecules are better able to withstand exonuclease attacks and exhibit increased stability compared to conventional antisense oligonucleotides.

[0463] In another preferred embodiment, the antisense oligodeoxynucleotides against herpes simplex virus types 1 and 2 described in International Application No. WO 95/04141 are used.

[0464] In yet another preferred embodiment, the covalently cross-linked antisense oligonucleotides described in International Application No. WO 96/31523 are used. These double- or single-stranded oligonucleotides comprise one or more, respectively, inter- or intra-oligonucleotide covalent cross-linkages, wherein the linkage consists of an amide bond between a primary amine group of one strand and a carboxyl group of the other strand or of the same strand, respectively, the primary amine group being directly substituted in the 2' position of the strand nucleotide monosaccharide ring, and the carboxyl group being carried by an aliphatic spacer group substituted on a nucleotide or nucleotide analog of the other strand or the same strand, respectively.

[0465] The antisense oligodeoxynucleotides and oligonucleotides disclosed in International Application No. WO 92/18522 may also be used. These molecules are stable to degradation and contain at least one transcription control recognition sequence which binds to control proteins and are effective as decoys therefor. These molecules may contain "hairpin" structures, "dumbbell" structures, "modified dumbbell" structures, "cross-linked" decoy structures and "loop" structures.

[0466] In another preferred embodiment, the cyclic double-stranded oligonucleotides described in European Patent Application No. 0 572 287 A2. These ligated oligonucleotide "dumbbells" contain the binding site for a transcription factor and inhibit expression of the gene under control of the transcription factor by sequestering the factor.

[0467] Use of the closed antisense oligonucleotides disclosed in International Application No. WO 92/19732 is also contemplated. Because these molecules have no free ends, they are more resistant to degradation by exonucleases than are conventional oligonucleotides. These oligonucleotides may be multifunctional, interacting with several regions which are not adjacent to the target mRNA.

[0468] The appropriate level of antisense nucleic acids required to inhibit gene expression may be determined using *in vitro* expression analysis. The antisense molecule may be introduced into the cells by diffusion, injection, infection or transfection using procedures known in the art. For example, the antisense nucleic acids can be introduced into the body as a bare or naked oligonucleotide, oligonucleotide encapsulated in lipid, oligonucleotide sequence encapsidated by viral protein, or as an oligonucleotide operably linked to a promoter contained in an expression vector. The expression vector may be any of a variety of expression vectors known in the art, including retroviral or viral vectors, vectors capable of extrachromosomal replication, or integrating vectors. The vectors may be DNA or RNA.

[0469] The antisense molecules are introduced onto cell samples at a number of different concentrations preferably between $1x10^{-10}$àM to $1x10^{-4}$àM. Once the minimum concentration that can adequately control gene expression is identified, the optimized dose is translated into a dosage suitable for use *in vivo*. For example, an inhibiting concentration in culture of $1x10^{-7}$à translates into a dose of approximately 0.6 mg/kg bodyweight. Levels of oligonucleotide approaching 100 mg/kg bodyweight or higher maybe possible after testing the toxicity of the oligonucleotide in laboratory animals. It is additionally contemplated that cells from the vertebrate are removed, treated with the antisense oligonucleotide, and reintroduced into the vertebrate.

[0470] It is further contemplated that the antisense oligonucleotide sequence is incorporated into a ribozyme sequence to enable the antisense to specifically bind and cleave its target mRNA. For technical applications of ribozyme and antisense oligonucleotides see Rossi *et al., supra*.

[0471] In a preferred application of this invention, the polypeptide encoded by the gene is first identified, so that the effectiveness of antisense inhibition on translation can be monitored using techniques that include but are not limited to antibody-mediated tests such as RIAs and ELISA, functional assays, or radiolabeling.

[0472] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may also be used in gene therapy approaches based on intracellular triple helix formation. Triple helix oligonucleotides are used to inhibit transcription from a genome. They are particularly useful for studying alterations in cell activity as it is associated with a particular gene. The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-

related nucleic acids of the present invention or, more preferably, a portion of those sequences, can be used to inhibit gene expression in individuals having diseases associated with expression of a particular gene. Similarly, the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids can be used to study the effect of inhibiting transcription of a particular gene within a cell. Traditionally, homopurine sequences were considered the most useful for triple helix strategies. However, homopyrimidine sequences can also inhibit gene expression. Such homopyrimidine oligonucleotides bind to the major groove at homopurine:homopyrimidine sequences. Thus, both types of sequences from the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are contemplated within the scope of this invention.

EXAMPLE 57

Preparation and use of Triple Helix Probes

[0473] The sequences of the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are scanned to identify 10-mer to 20-mer homopyrimidine or homopurine stretches which could be used in triple-helix based strategies for inhibiting gene expression. Following identification of candidate homopyrimidine or homopurine stretches, their efficiency in inhibiting gene expression is assessed by introducing varying amounts of oligonucleotides containing the candidate sequences into tissue culture cells which normally express the target gene. The oligonucleotides may be prepared on an oligonucleotide synthesizer or they may be purchased commercially from a company specializing in custom oligonucleotide synthesis, such as GENSET, Paris, France.

[0474] The oligonucleotides may be introduced into the cells using a variety of methods known to those skilled in the art, including but not limited to calcium phosphate precipitation, DEAE-Dextran, electroporation, liposome-mediated transfection or native uptake.

[0475] Treated cells are monitored for altered cell function or reduced gene expression using techniques such as Northern blotting, RNase protection assays, or PCR based strategies to monitor the transcription levels of the target gene in cells which have been treated with the oligonucleotide. The cell functions to be monitored are predicted based upon the homologies of the target genes corresponding to the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids from which the oligonucleotide were derived with known gene sequences that have been associated with a particular function. The cell functions can also be predicted based on the presence of abnormal physiologies within cells derived from individuals with a particular inherited disease, particularly when the EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids are associated with the disease using techniques described herein.

[0476] The oligonucleotides which are effective in inhibiting gene expression in tissue culture cells may then be introduced *in vivo* using the techniques described above and in Example 56 at a dosage calculated based on the *in vitro* results, as described in Example 56.

[0477] In some embodiments, the natural (beta) anomers of the oligonucleotide units can be replaced with alpha anomers to render the oligonucleotide more resistant to nucleases. Further, an intercalating agent such as ethidium bromide, or the like, can be attached to the 3' end of the alpha oligonucleotide to stabilize the triple helix. For information on the generation of oligonucleotides suitable for triple helix formation see Griffin *et al.* (*Science* **245**:967-971 (1989)).

EXAMPLE 58

Use of EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids to express an Encoded Protein in a Host Organism

[0478] The EST-related nucleic acids, positional segments of EST-related nucleic acids or fragments of positional segments of EST-related nucleic acids may also be used to express an encoded protein or polypeptide in a host organism to produce a beneficial effect. In addition, nucleic acids encoding the EST-related polypeptides, positional segments of EST-related polypeptides or fragments of positional segments of EST-related polypeptides may be used to express the encoded protein or polypeptide in a host organism to produce a beneficial effect.

[0479] In such procedures, the encoded protein or polypeptide may be transiently expressed in the host organism or stably expressed in the host organism. The encoded protein or polypeptide may have any of the activities described above. The encoded protein or polypeptide may be a protein or polypeptide which the host organism lacks or, alternatively, the encoded protein may augment the existing levels of the protein in the host organism.

[0480] In some embodiments in which the protein or polypeptide is secreted, nucleic acids encoding the full length protein (i.e. the signal peptide and the mature protein), or nucleic acids encoding only the mature protein (i.e. the protein generated when the signal peptide is cleaved off) is introduced into the host organism.

[0481] The nucleic acids encoding the proteins or polypeptides may be introduced into the host organism using a variety of techniques known to those of skill in the art. For example, the extended cDNA may be injected into the host organism as naked DNA such that the encoded protein is expressed in the host organism, thereby producing a beneficial effect.

[0482] Alternatively, the nucleic acids encoding the protein or polypeptide may be cloned into an expression vector downstream of a promoter which is active in the host organism. The expression vector may be any of the expression vectors designed for use in gene therapy, including viral or retroviral vectors. The expression vector may be directly introduced into the host organism such that the encoded protein is expressed in the host organism to produce a beneficial effect. In another approach, the expression vector may be introduced into cells *in vitro*. Cells containing the expression vector are thereafter selected and introduced into the host organism, where they express the encoded protein or polypeptide to produce a beneficial effect.

EXAMPLE 59

Use of Signal Peptides To Import Proteins Into Cells

[0483] The short core hydrophobic region (h) of signal peptides encoded by the sequences of SEQ ID NOs: 24-652 and 3721-3811 may also be used as a carrier to import a peptide or a protein of interest, so-called cargo, into tissue culture cells (Lin*et al., J. Biol. Chem.*, 270: 14225-14258 (1995); Du *et al., J. Peptide Res.*, **51**: 235-243 (1998); Rojas *et al., Nature Biotech.*, **16**: 370-375 (1998)).

[0484] When cell permeable peptides of limited size (approximately up to 25 amino acids) are to be translocated across cell membrane, chemical synthesis may be used in order to add the h region to either the C-terminus or the N-terminus to the cargo peptide of interest. Alternatively, when longer peptides or proteins are to be imported into cells, nucleic acids can be genetically engineered, using techniques familiar to those skilled in the art, in order to link the extended cDNA sequence encoding the h region to the 5' or the 3' end of a DNA sequence coding for a cargo polypeptide. Such genetically engineered nucleic acids are then translated either *in vitro* or *in vivo* after transfection into appropriate cells, using conventional techniques to produce the resulting cell permeable polypeptide. Suitable hosts cells are then simply incubated with the cell permeable polypeptide which is then translocated across the membrane.

[0485] This method may be applied to study diverse intracellular functions and cellular processes. For instance, it has been used to probe functionally relevant domains of intracellular proteins and to examine protein-protein interactions involved in signal transduction pathways (Lin *et al. supra;* Lin *et al., J. Biol. Chem.*, 271: 5305-5308 (1996); Rojas *et al., J. Biol. Chem.*, 271: 27456-27461 (1996); Liu *et al., Proc. Natl, Acad. Sci. USA*, 93: 11819-11824 (1996); Rojas *et al., Bioch. Biophys. Res. Commun.*, 234: 675-680 (1997)).

[0486] Such techniques may be used in cellular therapy to import proteins producing therapeutic effects. For instance, cells isolated from a patient may be treated with imported therapeutic proteins and then re-introduced into the host organism.

[0487] Alternatively, the h region of signal peptides of the present invention could be used in combination with a nuclear localization signal to deliver nucleic acids into cell nucleus. Such oligonucleotides may be antisense oligonucleotides or oligonucleotides designed to form triple helixes, as describedabove, in order to inhibit processing and maturation of a target cellular RNA.

EXAMPLE 60

Computer Embodiments

[0488] As used herein the term "nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681" encompasses the nucleotide sequences of SEQ ID NOs: 24-4100 and 8178-36681, fragments of SEQ ID NOs: 24-4100 and 8178-36681, nucleotide sequences homologous to SEQ ID NOs: 24-4100 and 8178-36681 or homologous to fragments of SEQ ID NOs: 24-4100 and 8178-36681, and sequences complementary to all of the preceding sequences. The fragments include portions of SEQ ID NOs: 24-4100 and 8178-36681 comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive nucleotides of SEQ ID NOs: 24-4100 and 8178-36681. Preferably, the fragments are novel fragments. Homologous sequences and fragments of SEQ ID NOs: 24-4100 and 8178-36681 refer to a sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, or 75% homology to these sequences. Homology may be determined using any of the computer programs and parameters described in Example 18, including BLAST2N with the default parameters or with any modified parameters. Homologous sequences also include RNA sequences in which uridines replace the thymines in the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error as described above. It will be appreciated that the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 can be represented in the traditional single character format (See the inside back cover of Starrier, Lubert. *Biochemistry,* 3rdà edition. W. H Freeman & Co., New York.) or in any other format which records the identity of the nucleotides in a sequence.

[0489] As used herein the term "polypeptide codes of SEQ ID NOs: 4101-8177" encompasses the polypeptide sequence of SEQ ID NOs: 4101-8177 which are encoded by the 5' EST s of SEQ ID NOs: 24-4100 and 8178-36681, polypeptide sequences homologous to the polypeptides of SEQ ID NOs: 4101-8177, or fragments of any of the preceding sequences. Homologous polypeptide sequences refer to a polypeptide sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75% homology to one of the polypeptide sequences of SEQ ID NOs: 4101-8177. Homology may be determined using any of the computer programs and parameters described herein, including FASTA with the default parameters or with any modified parameters. The homologous sequences may be obtained using any of

the procedures described herein or may result from the correction of a sequencing error as described above. The polypeptide fragments comprise at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of the polypeptides of SEQ ID NOs: 4101-8177. Preferably, the fragments are novel fragments. It will be appreciated that the polypeptide codes of the SEQ ID NOs: 4101-8177 can be represented in the traditional single character format or three letter format (See the inside back cover of Starrier, Lubert. *Biochemistry,* 3rd à edition. W. H Freeman & Co., New York.) or in any other format which relates the identity of the polypeptides in a sequence.

[0490] It will be appreciated by those skilled in the art that the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 and polypeptide codes of SEQ ID NOs: 4101-8177 can be stored, recorded, and manipulated on any medium which can be read and accessed by a computer. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681, one or more of the polypeptide codes of SEQ ID NOs: 4101-8177. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 polypeptide codes of SEQ ID NOs: 4101-8177.

[0491] Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disc, a floppy disc, a magnetic tape, CD-ROM, DVD, RAM, or ROM as well as other types of other media known to those skilled in the art.

[0492] Embodiments of the present invention include systems, particularly computer systems which contain the sequence information described herein. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze the nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681, or the amino acid sequences of the polypeptide codes of SEQ ID NOs: 4101-8177. The computer system preferably includes the computer readable media described above, and a processor for accessing and manipulating the sequence data.

[0493] Preferably, the computer is a general purpose system that comprises a central processing unit (CPU), one or more data storage components for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

[0494] In one particular embodiment, the computer system includes a processor connected to a bus which is connected to a main memory (preferably implemented as RAM) and one or more data storage devices, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system further includes one or more data retrieving devices for reading the data stored on the data storage components. The data retrieving device may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, etc. In some embodiments, the data storage component is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device. Software for accessing and processing the nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681, or the amino acid sequences of the polypeptide codes of SEQ ID NOs: 4101-8177 (such as search tools, compare tools, and modeling tools etc.) may reside in main memory during execution.

[0495] In some embodiments, the computer system may further comprise a sequence comparer for comparing the above-described nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or polypeptide codes of SEQ ID NOs: 4101-8177 stored on a computer readable medium to reference nucleotide or polypeptide sequences stored on a computer readable medium. A "sequence comparer" refers to one or more programs which are implemented on the computer system to compare a nucleotide or polypeptide sequence with other nucleotide or polypeptide sequences and/or compounds including but not limited to peptides, peptidomimetics, and chemicals stored within the data storage means. For example, the sequence comparer may compare the nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681, or the amino acid sequences of the polypeptide codes of SEQ ID NOs: 4101-8177 stored on a computer readable medium to reference sequences stored on a computer readable medium to identify homologies, motifs implicated in biological function, or structural motifs. The various sequence comparer programs identified elsewhere in this patent specification are particularly contemplated for use in this aspect of the invention.

[0496] Accordingly, one aspect of the present invention is a computer system comprising a processor, a data storage device having stored thereon a nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 or a polypeptide code of SEQ ID NOs: 4101-8177, a data storage device having retrievably stored thereon reference nucleotide sequences or polypeptide sequences to be compared to the nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 or polypeptide code of SEQ ID NOs: 4101-8177 and a sequence comparer for conducting the comparison. The sequence comparer may indicate a homology level between the sequences compared or identify structural motifs in the above described nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 and polypeptide codes of SEQ ID NOs: 4101-8177 or it may identify structural motifs in sequences which are compared to these nucleic acid codes and polypeptide codes. In some embodiments, the data storage device may have stored thereon the sequences of at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or polypeptide codes of SEQ ID NOs: 4101-8177.

[0497] Another aspect of the present invention is a method for determining the level of homology between a nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 and a reference nucleotide sequence, comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through the use of a computer program which determines homology levels and determining homology between the nucleic acid code and the reference nucleotide sequence with the computer program. The computer program may be any of a number of computer programs for determining homology levels, including those specifically enumerated herein, including BLAST2N with the default parameters or with any modified parameters. The method may be implemented using the computer systems described above. The method may also be performed by reading 2, 5, 10, 15, 20, 25, 30, or 50 of the above described nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 through use of the computer program and determining homology between the nucleic acid codes and reference nucleotide sequences.

[0498] Alternatively, the computer program may be a computer program which compares the nucleotide sequences of the nucleic acid codes of the present invention, to reference nucleotide sequences in order to determine whether the nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 differs from a reference nucleic acid sequence at one or more positions. Optionally such a program records the length and identity of inserted, deleted or substituted nucleotides with respect to the sequence of either the reference polynucleotide or the nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681. In one embodiment, the computer program may be a program which determines whether the nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 contain a single nucleotide polymorphism (SNP) with respect to a reference nucleotide sequence. This single nucleotide polymorphism may comprise a single base substitution, insertion, or deletion.

[0499] Another aspect of the present invention is a method for determining the level of homology between a polypeptide code of SEQ ID NOs: 4101-8177 and a reference polypeptide sequence, comprising the steps of reading the polypeptide code of SEQ ID NOs: 4101-8177 and the reference polypeptide sequence through use of a computer program which determines homology levels and determining homology between the polypeptide code and the reference polypeptide sequence using the computer program.

[0500] Accordingly, another aspect of the present invention is a method for determining whether a nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 differs at one or more nucleotides from a reference nucleotide sequence comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through use of a computer program which identifies differences between nucleic acid sequences and identifying differences between the nucleic acid code and the reference nucleotide sequence with the computer program. In some embodiments, the computer program is a program which identifies single nucleotide polymorphisms. The method may be implemented by the computer systems described above. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 and the reference nucleotide sequences through the use of the computer program and identifying differences between the nucleic acid codes and the reference nucleotide sequences with the computer program.

[0501] In other embodiments the computer based system may further comprise an identifier for identifying features within the nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the amino acid sequences of the polypeptide codes of SEQ ID NOs: 4101-8177.

[0502] An "identifier" refers to one or more programs which identifies certain features within the above-described nucleotide sequences of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the amino acid sequences of the polypeptide codes of SEQ ID NOs: 4101-8177. In one embodiment, the identifier may comprise a program which identifies an open reading frame in the cDNAs codes of SEQ ID NOs: 24-4100 and 8178-36681.

[0503] In another embodiment, the identifier may comprise a molecular modeling program which determines the 3-dimensional structure of the polypeptides codes of SEQ ID NOs: 4101-8177. In some embodiments, the molecular modeling program identifies target sequences that are most compatible with profiles representing the structural environments of the residues in known three-dimensional protein structures. (See, e.g., Eisenberg et al., U.S. Patent No. 5,436,850 issued July 25, 1995). In another technique, the known three-dimensional structures of proteins in a given family are superimposed to define the structurally conserved regions in that family. This protein modeling technique also uses the known three-dimensional structure of a homologous protein to approximate the structure of the polypeptide codes of SEQ ID NOs: 4101-8177. (See e.g., Srinivasan, et al., U.S. Patent No. 5,557,535 issued September 17, 1996). Conventional homology modeling techniques have been used routinely to build models of proteases and antibodies. (Sowdhamini et al., Protein Engineering 10:207, 215 (1997)). Comparative approaches can also be used to develop three-dimensional protein models when the protein of interest has poor sequence identity to template proteins. In some cases, proteins fold into similar three-dimensional structures despite having very weak sequence identities. For example, the three-dimensional structures of a number of helical cytokines fold in similar three-dimensional topology in spite of weak sequence homology.

[0504] The recent development of threading methods now enables the identification of likely folding patterns in a number of situations where the structural relatedness between target and template(s) is not detectable at the sequence level. Hybrid methods, in which fold recognition is performed using Multiple Sequence Threading (MST), structural equivalencies are deduced from the threading output using a distance geometry program DRAGON to construct a low resolution model, and a full-atom representation is constructed using a molecular modeling package such as QUANTA.

[0505] According to this 3-step approach, candidate templates are first identified by using the novel fold recognition algorithm MST, which is capable of performing simultaneous threading of multiple aligned sequences onto one or more 3-D structures. In a second step, the structural equivalencies obtained from the MST output are converted into interresidue distance restraints and fed into the distance geometry program DRAGON, together with

auxiliary information obtained from secondary structure predictions. The program combines the restraints in an unbiased manner and rapidly generates a large number of low resolution model confirmations. In a third step, these low resolution model confirmations are converted into full-atom models and subjected to energy minimization using the molecular modeling package QUANTA. (See e.g., Aszódi et al., Proteins:Structure, Function, and Genetics, Supplement 1:38-42 (1997)).

[0506] The results of the molecular modeling analysis may then be used in rational drug design techniques to identify agents which modulate the activity of the polypeptide codes of SEQ ID NOs: 4101-8177.

[0507] Accordingly, another aspect of the present invention is a method of identifying a feature within the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177 comprising reading the nucleic acid code(s) or the polypeptide code(s) through the use of a computer program which identifies features therein and identifying features within the nucleic acid code(s) or polypeptide code(s) with the computer program. In one embodiment, computer program comprises a computer program which identifies open reading frames. In a further embodiment, the computer program identifies structural motifs in a polypeptide sequence. In another embodiment, the computer program comprises a molecular modeling program. The method may be performed by reading a single sequence or at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177 through the use of the computer program and identifying features within the nucleic acid codes or polypeptide codes with the computer program.

[0508] The nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177 may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177 may be stored as text in a word processing file, such as MicrosoftWORD or WORDPERFECT or as an ASCII file in a variety of database programs familiar to those of skill in the art, such as DB2, SYBASE, or ORACLE. In addition, many computer programs and databases may be used as sequence comparers, identifiers, or sources of reference nucleotide or polypeptide sequences to be compared to the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177. The following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid codes of SEQ ID NOs: 24-4100 and 8178-36681 or the polypeptide codes of SEQ ID NOs: 4101-8177. The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul et al, *J. Mol. Biol.* 215: 403 (1990)), FASTA (Pearson and Lipman, *Proc. Natl. Acad. Sci. USA,* 85: 2444 (1988)), FASTDB (Brutlag et al. Comp. App. Biosci. 6:237-245, 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius^{2}à.DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.), Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab Diversity Explorer (Molecular Simulations Inc.), Gene Explorer (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the EMBL/Swissprotein database, the MDL Available Chemicals Directory database, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseqn database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

[0509] Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

EXAMPLE 61

Methods of Making Nucleic Acids

[0510] The present invention also comprises methods of making the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of the EST-related nucleic acids, or fragments of positional segments of the EST-related nucleic acids. The methods comprise sequentially linking together nucleotides to produce the nucleic acids having the preceding sequences. A variety of methods of synthesizing nucleic acids are known to those skilled in the art.

[0511] In many of these methods, synthesis is conducted on a solid support. These included the 3' phosphoramidite methods in which the 3' terminal base of the desired oligonucleotide is immobilized on an insoluble carrier. The nucleotide base to be added is blocked at the 5' hydroxyl and activated at the 3' hydroxyl so as to cause coupling with the immobilized nucleotide base. Deblocking of the new immobilized nucleotide compound and repetition of the cycle will produce the desired polynucleotide. Alternatively, polynucleotides may be prepared as described in U.S. Patent No. 5,049,656. In some embodiments, several polynucleotides prepared as described above are ligated together to generate longer polynucleotides having a desired sequence.

EXAMPLE 62

Methods of Making Polypeptides

[0512] The present invention also comprises methods of making the polynucleotides encoded by EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of the EST-related nucleic acids, or fragments of positional segments of the EST-related nucleic acids and methods of making the EST-related polypeptides, fragments of EST-related polypeptides, positional segments of EST-related polypeptides, or fragments of EST-related polypeptides. The methods comprise sequentially linking together amino acids to produce the nucleic polypeptides having the preceding sequences. In some embodiments, the polypeptides made by these methods are 150 amino acid or less in length. In other embodiments, the polypeptides made by these methods are 120 amino acids or less in length.

[0513] A variety of methods of making polypeptides are known to those skilled in the art, including methods in which the carboxyl terminal amino acid is bound to polyvinyl benzene or another suitable resin. The amino acid to be added possesses blocking groups on its amino moiety and any side chain reactive groups so that only its carboxyl moiety can react. The carboxyl group is activated with carbodiimide or another activating agent and allowed to couple to the immobilized amino acid. After removal of the blocking group, the cycle is repeated to generate a polypeptide having the desired sequence. Alternatively, the methods described in U.S. Patent No. 5,049,656 may be used.

[0514] As discussed above, the EST-related nucleic acids, fragments of the EST-related nucleic acids, positional segments of the EST-related nucleic acids, or fragments of positional segments of the EST-related nucleic acids can be used for various purposes. The polynucleotides can be used to express recombinant protein for analysis, characterization or therapeutic use; production of secreted polypeptides or chimeric polypeptides, antibody production, as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in disease states); as molecular weight markers on Southern gels; as chromosome markers or tags (when labeled) to identify chromosomes or to map related gene positions; to compare with endogenous DNA sequences in patients to identify potential genetic disorders; as probes to hybridize and thus discover novel, related DNA sequences; as a source of information to derive PCR primers for genetic fingerprinting; for selecting and making oligomers for attachment to a "gene chip" or other support, including for examination for expression patterns; to raise anti-protein antibodies using DNA immunization techniques; and as an antigen to raise anti-DNA antibodies or elicit another immune response. Where the polynucleotide encodes a protein or polypeptide which binds or potentially binds to another protein or polypeptide (such as, for example, in a receptor-ligand interaction), the polynucleotide can also be used in interaction trap assays (such as, for example, that described in Gyuris *et al.*, *Cell***75**:791-803 (1993)) to identify polynucleotides encoding the other protein or polypeptide with which binding occurs or to identify inhibitors of the binding interaction.

[0515] The proteins or polypeptides provided by the present invention can similarly be used in assays to determine biological activity, including in a panel of multiple proteins for high-throughput screening; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its receptor) in biological fluids; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); and, of course, to isolate correlative receptors or ligands. Where the protein or polypeptide binds or potentially binds to another protein or polypeptide (such as, for example, in a receptor-ligand interaction), the protein can be used to identify the other protein with which binding occurs or to identify inhibitors of the binding interaction. Proteins or polypeptides involved in these binding interactions can also be used to screen for peptide or small molecule inhibitors or agonists of the binding interaction.

[0516] Any or all of these research utilities are capable of being developed into reagent grade or kit format for commercialization as research products.

[0517] Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include without limitation "Molecular Cloning; A Laboratory Manual", 2d ed., Cold Spring Harbor Laboratory Press, Sambrook, J., E.F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology; Guide to Molecular Cloning Techniques", Academic Press, Berger, S.L. and A.R. Kimmel eds., 1987.

[0518] Polynucleotides and proteins or polypeptides of the present invention can also be used as nutritional sources or supplements. Such uses include without limitation use as a protein or amino acid supplement, use as a carbon source, use as a nitrogen source and use as a source of carbohydrate. In such cases the protein or polynucleotide of the invention can be added to the feed of a particular organism or can be administered as a separate solid or liquid preparation, such as in the form of powder, pills, solutions, suspensions or capsules. In the case of microorganisms, the protein or polynucleotide of the invention can be added to the medium in or on which the microorganism is cultured.

[0519] Although this invention has been described in terms of certain preferred embodiments, other embodiments which will be apparent to those of ordinary skill in the art in view of the disclosure herein are also within the scope of this invention. Accordingly, the scope of the invention is intended to be defined only by reference to the appended claims

Claims

1. A purified nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and

SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

2. A purified nucleic acid comprising at least 10 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

3. A purified nucleic acid comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

4. A purified nucleic acid comprising the coding sequence of a sequence selected from the group consisting of SEQ ID NOs: 24-4100.

5. A purified nucleic acid comprising the full coding sequences of a sequence selected from the group consisting of SEQ ID NOs: 3721-3811 wherein the full coding sequence comprises the sequence encoding the signal peptide and the sequence encoding the mature protein.

6. A purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs: 3721-3811 which encodes the mature protein.

7. A purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs: 24-652 and 3721-3811 which encode the signal peptide.

8. A purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

9. A purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 7798-7888.

10. A purified nucleic acid encoding a polypeptide comprising a mature protein included in a sequence selected from the group consisting of the sequences of SEQ ID NOs: 7798-7888.

11. A purified nucleic acid encoding a polypeptide comprising a signal peptide included in a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-4729 and 7798-7888.

12. A purified nucleic acid at least 15 nucleotides in length which hybridizes under stringent conditions to a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

13. A purified or isolated polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

14. A purified or isolated polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 7798-7888.

15. A purified or isolated polypeptide comprising a mature protein of a polypeptide selected from the group consisting of SEQ ID NOs: 7798-7888.

16. A purified or isolated polypeptide comprising a signal peptide of a sequence selected from the group consisting of the polypeptides of SEQ ID NOs: 4101-4729 and 7798-7888.

17. A purified or isolated polypeptide comprising at least 10 consecutive amino acids of a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4101-8177.

18. A method of making a cDNA comprising the steps of:

contacting a collection of mRNA molecules from human cells with a primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681;

hybridizing said primer to an mRNA in said collection that encodes said protein;

reverse transcribing said hybridized primer to make a first cDNA strand from said mRNA;

making a second cDNA strand complementary to said first cDNA strand; and

isolating the resulting cDNA encoding said protein comprising said first cDNA strand and said second cDNA strand.

19. A purified cDNA obtainable by the method of Claim 18.

20. The cDNA of Claim 19 wherein said cDNA encodes at least a portion of a human polypeptide.

21. A method of making a cDNA comprising the steps of:

obtaining a cDNA comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681;

contacting said cDNA with a detectable probe comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 under conditions which permit said probe to hybridize to said cDNA;

identifying a cDNA which hybridizes to said detectable probe; and

isolating said cDNA which hybridizes to said probe.

22. A purified cDNA obtainable by the method of Claim 21.

23. The cDNA of Claim 22 wherein said cDNA encodes at least a portion of a human polypeptide.

24. A method of making a cDNA comprising the steps of:

contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of said mRNA;

hybridizing said first primer to said polyA tail;

reverse transcribing said mRNA to make a first cDNA strand;

making a second cDNA strand complementary to said first cDNA strand using at least one primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681; and

isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

25. A purified cDNA obtainable by the method of Claim 24.

26. The cDNA of Claim 25 wherein said cDNA encodes at least a portion of a human polypeptide.

27. The method of Claim 24, wherein the second cDNA strand is made by:

contacting said first cDNA strand with a first pair of primers, said first pair of primers comprising a second primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and a third primer having a sequence therein which is included within the sequence of said first primer;

performing a first polymerase chain reaction with said first pair of primers to generate a first PCR product;

contacting said first PCR product with a second pair of primers, said second pair of primers comprising a fourth primer, said fourth primer comprising at least 15 consecutive nucleotides of said sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, and a fifth primer, wherein said fourth and fifth hybridize to sequences within said first PCR product; and

performing a second polymerase chain reaction, thereby generating a second PCR product.

28. A purified cDNA obtainable by the method of Claim 27.

29. The cDNA of Claim 28 wherein said cDNA encodes at least a portion of a human polypeptide.

30. The method of Claim 24 wherein the second cDNA strand is made by:

contacting said first cDNA strand with a second primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681;

hybridizing said second primer to said first strand cDNA; and

extending said hybridized second primer to generate said second cDNA strand.

31. A purified cDNA obtainable by the method of Claim 30.

32. The cDNA of Claim 28, wherein said cDNA encodes at least a portion of a human polypeptide.

33. A method of making a polypeptide comprising the steps of:

obtaining a cDNA which encodes a polypeptide encoded by a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 or a cDNA which encodes a polypeptide comprising at least 10 consecutive amino acids of a polypeptide encoded by a sequence selected from the group consisting of SEQ ID NOs: 24-4100;

inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter;

introducing said expression vector into a host cell whereby said host cell produces the protein encoded by said cDNA; and

isolating said protein.

34. An isolated protein obtainable by the method of Claim 33.

35. A method of obtaining a promoter DNA comprising the steps of:

obtaining genomic DNA located upstream of a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681;

screening said genomic DNA to identify a promoter capable of directing transcription initiation; and

isolating said DNA comprising said identified promoter.

36. The method of Claim 35, wherein said obtaining step comprises walking from genomic DNA comprising a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

37. The method of Claim 36, wherein said screening step comprises inserting genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs. 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 into a promoter reporter vector.

38. The method of Claim 36, wherein said screening step comprises identifying motifs in genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 which are transcription factor binding sites or transcription start sites.

39. An isolated promoter obtainable by the method of any one of Claims 34 to 38.

40. In an array of discrete ESTs or fragments thereof of at least 15 nucleotides in length, the improvement comprising inclusion in said array of at least one sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and fragments comprising at least 15 consecutive nucleotides of said sequence.

41. The array of Claim 40 including therein at least two sequences selected from the group consisting of SEQ ID NOs:

24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, and fragments comprising at least 15 consecutive nucleotides of said sequences.

42. The array of Claim 40 including therein at least five sequences selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681, the sequences complementary to the sequences of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and fragments comprising at least 15 consecutive nucleotides of said sequences.

43. An enriched population of recombinant nucleic acids, said recombinant nucleic acids comprising an insert nucleic acid and a backbone nucleic acid, wherein at least 5% of said insert nucleic acids in said population comprise a sequence selected from the group consisting of SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681 and the sequences complementary to SEQ ID NOs: 24-4100 and SEQ ID NOs: 8178-36681.

44. A purified or isolated antibody capable of specifically binding to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 4101-8177.

45. A purified or isolated antibody capable of specifically binding to a polypeptide comprising at least 10 consecutive amino acids of a sequence selected from the group consisting of SEQ ID NOs: 4101-8177.

46. An antibody composition capable of selectively binding to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 8 amino acids of any of SEQ ID NOs: 4101-8177, wherein said antibody is polyclonal or monoclonal.

47. A computer readable medium having stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQ ID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177.

48. A computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177.

49. The computer system of Claim 48 further comprising a sequence comparer and a data storage device having reference sequences stored thereon.

50. The computer system of Claim 49 wherein said sequence comparer comprises a computer program which indicates polymorphisms.

51. The computer system of Claim 48 further comprising an identifier which identifies features in said sequence.

52. A method for comparing a first sequence to a reference sequence wherein said first sequence is selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177 comprising the steps of:

reading said first sequence and said reference sequence through use of a computer program which compares sequences; and

determining differences between said first sequence and said reference sequence with said computer program.

53. The method of Claim 52, wherein said step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

54. A method for identifying a feature in a sequence selected from the group consisting of a nucleic acid code of SEQID NOs: 24-4100 and 8178-36681 and a polypeptide code of SEQ ID NOs: 4101-8177 comprising the steps of:

reading said sequence through the use of a computer program which identifies features in sequences; and

identifying features in said sequence with said computer program.

55. A vector comprising a nucleic acid according to any one of Claims 1 to 12.

56. A host cell containing a nucleic acid of Claim 55.

57. A method of making a nucleic acid of Claims 1 comprising the steps of:

introducing said nucleic acid into a host cell such that said nucleic acid is present in multiple copies in

each host cell; and

isolating said nucleic acid from said host cell.

58. A method of making a nucleic acid of any one of Claims 1 to 12 comprising the step of sequentially linking together the nucleotides in said nucleic acids.

59. A method of making a polypeptide of any one of Claims 13 to 17 wherein said polypeptides is 150 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptides.

60. A method of making a polypeptide of any one of Claims 13 to 17 wherein said polypeptides is 120 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptides.

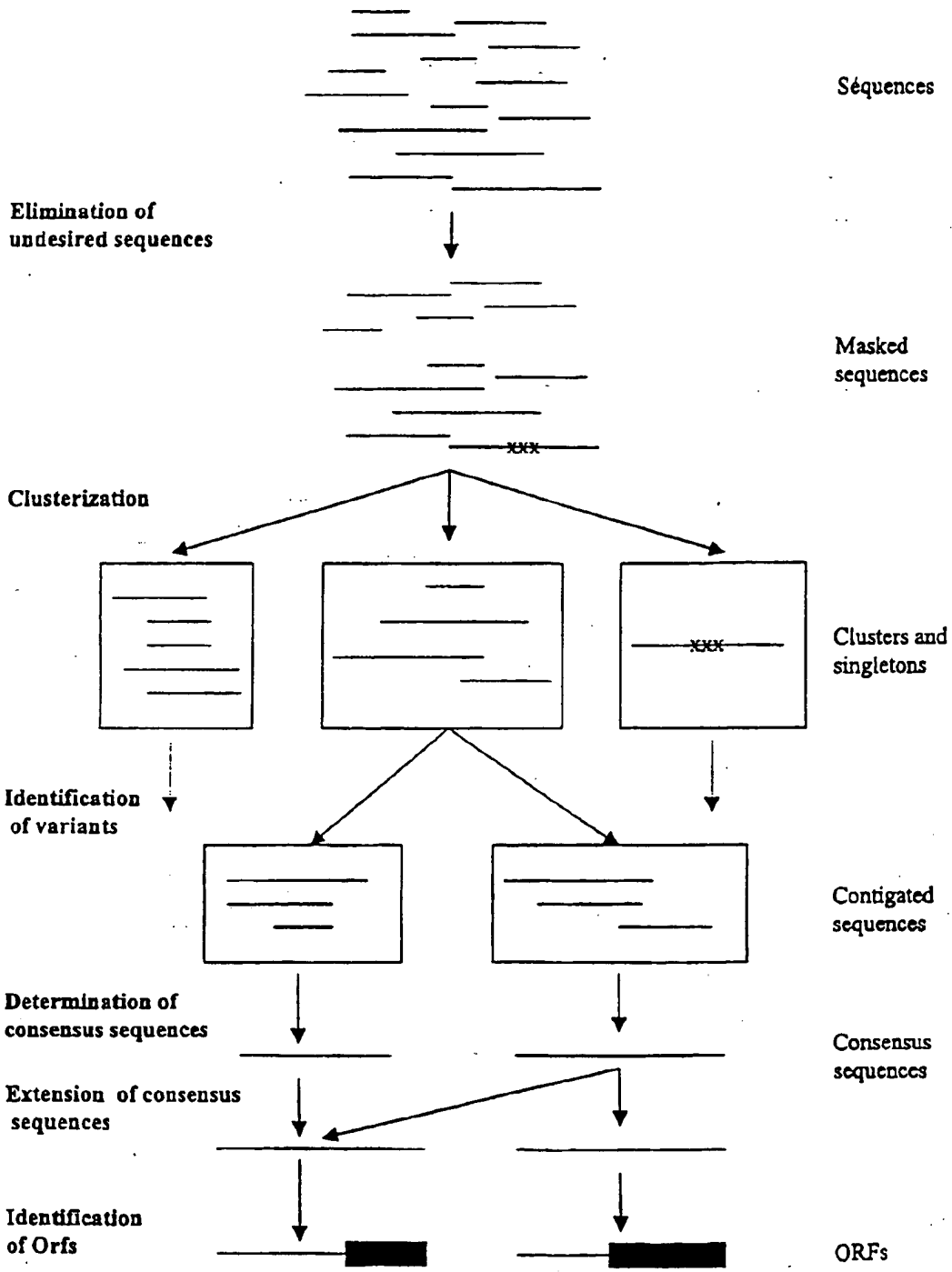
Séquences
Elimination of undesired sequences
Masked sequences
Clusterization
Clusters and singletons
Identification of variants
Contigated sequences
Determination of consensus sequences
Consensus sequences
Extension of consensus sequences
Identification of Orfs
ORFs

FIGURE 1

Minimum signal peptide score	false positive rate	false negative rate	proba(0.1)	proba(0.2)
3,5	0,121	0,036	0,467	0,664
4	0,096	0,06	0,519	0,708
4,5	0,078	0,079	0,565	0,745
5	0,062	0,098	0,615	0,782
5,5	0,05	0,127	0,659	0,813
6	0,04	0,163	0,694	0,836
6,5	0,033	0,202	0,725	0,855
7	0,025	0,248	0,763	0,878
7,5	0,021	0,304	0,78	0,889
8	0,015	0,368	0,816	0,909
8,5	0,012	0,418	0,836	0,92
9	0,009	0,512	0,856	0,93
9,5	0,007	0,581	0,863	0,934
10	0,006	0,679	0,835	0,919

FIGURE 2

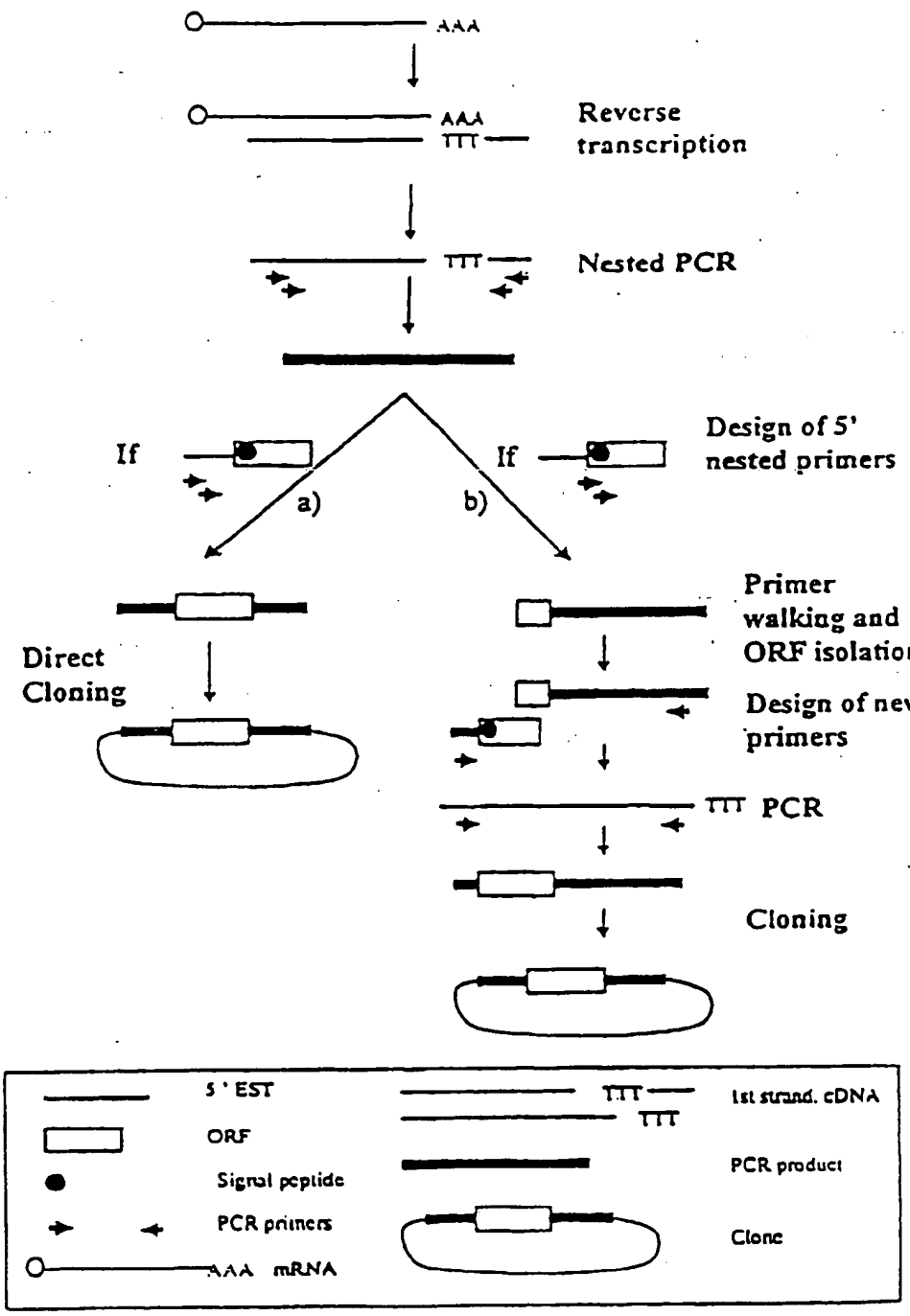
AAA
AAA
TTT
Reverse transcription
TTT
Nested PCR
If
a)
b)
If
Design of 5' nested primers
Direct Cloning
Primer walking and ORF isolation
Design of new primers
TTT PCR
Cloning
5' EST
ORF
Signal peptide
PCR primers
AAA mRNA
TTT
TTT
1st strand. cDNA
PCR product
Clone

FIGURE 3

Description of promoters structure isolated from SignalTag 5 'ESTs

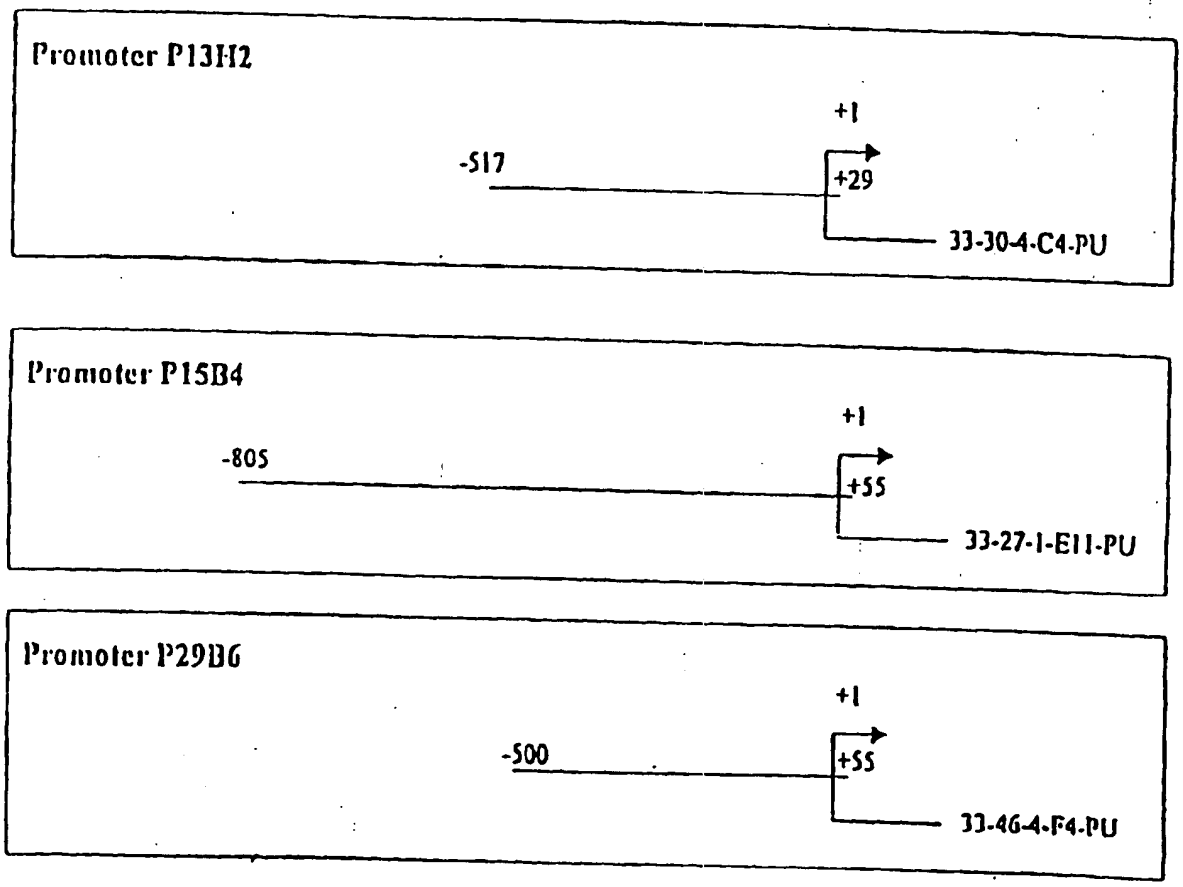

FIGURE 4

Description of Transcription Factor Binding Sites present on promoters isolated from SignalTag sequences

Promoter sequence P13H2 (546 bp):

Matrix	Position	Orientation	Score	Length	Sequence	Location in: SEQ ID NO: 17
CMYB_01	-502	+	0.983	9	TGTCAGTTG	17-25
MYOD_Q6	-501	-	0.961	10	CCCAACTGAC	complement of 18-27
S8_01	-444	-	0.960	11	AATAGAATTAG	complement of 75-85
S8_01	-425	+	0.966	11	AACTAAATTAG	94-104
DELTAEF1_01	-390	-	0.960	11	GCACACCTCAG	complement of 129-139
GATA_C	-364	-	0.964	11	AGATAAATCCA	complement of 155-165
CMYB_01	-349	+	0.958	9	CTTCAGTTG	170-178
GATA1_02	-343	+	0.959	14	TTGTAGATAGGACA	176-189
GATA_C	-339	+	0.953	11	AGATAGGACAT	180-190
TAL1ALPHAE47_01	-235	+	0.973	16	CATAACAGATGGTAAG	284-299
TAL1BETAE47_01	-235	+	0.983	16	CATAACAGATGGTAAG	284-299
TAL1BETAITF2_01	-235	+	0.978	16	CATAACAGATGGTAAG	284-299
MYOD_Q6	-232	-	0.954	10	ACCATCTGTT	complement of 287-296
GATA1_04	-217	-	0.953	13	TCAAGATAAAGTA	complement of 302-314
IK1_01	-126	+	0.963	13	AGTTGGGAATTCC	393-405
IK2_01	-126	+	0.985	12	AGTTGGGAATTC	393-404
CREL_01	-123	+	0.962	10	TGGGAATTCC	396-405
GATA1_02	-96	+	0.950	14	TCAGTGATATGGCA	423-436
SRY_02	-41	-	0.951	12	TAAAACAAAACA	complement of 478-489
E2F_02	-33	+	0.957	8	TTTAGCGC	486-493
MZF1_01	-5	-	0.975	8	TGAGGGGA	complement of 514-521

Promoter sequence P15B4 (861 bp):

Matrix	Position	Orientation	Score	Length	Sequence	Location in: SEQ ID NO: 20
NFY_Q6	-748	-	0.956	11	GGACCAATCAT	complement of 60-70
MZF1_01	-738	+	0.962	8	CCTGGGGA	70-77
CMYB_01	-684	+	0.994	9	TGACCGTTG	124-132
VMYB_02	-682	-	0.985	9	TCCAACGGT	complement of 126-134
STAT_01	-673	+	0.968	9	TTCCTGGAA	135-143
STAT_01	-673	-	0.951	9	TTCCAGGAA	complement of 135-143
MZF1_01	-556	-	0.956	8	TTGGGGGA	complement of 252-259
IK2_01	-451	+	0.965	12	GAATGGGATTTC	357-368
MZF1_01	-424	+	0.986	8	AGAGGGGA	384-391
SRY_02	-398	-	0.955	12	GAAAACAAAACA	complement of 410-421
MZF1_01	-216	+	0.960	8	GAAGGGGA	592-599
MYOD_Q6	-190	+	0.981	10	AGCATCTGCC	618-627
DELTAEF1_01	-176	-	0.958	11	TCCCACCTTCC	632-642
S8_01	5	-	0.992	11	GAGGCAATTAT	complement of 813-823
MZF1_01	16	-	0.986	8	AGAGGGGA	complement of 824-831

Promoter sequence P29B6 (555 bp):

Matrix	Position	Orientation	Score	Length	Sequence	Location in: SEQ ID NO: 23
ARNT_01	-311	-	0.964	16	GGACTCACGTGCTGCT	191-206
NMYC_01	-309	-	0.965	12	ACTCACGTGCTG	193-204
USF_01	-309	+	0.985	12	ACTCACGTGCTG	193-204
USF_01	-309	-	0.985	12	CAGCACGTGAGT	complement of 193-204
NMYC_01	-309	-	0.956	12	CAGCACGTGAGT	complement of 193-204
MYCMAX_02	-309	-	0.972	12	CAGCACGTGAGT	complement of 193-204
USF_C	-307	+	0.997	8	TCACGTGC	195-202
USF_C	-307	-	0.991	8	GCACGTGA	complement of 195-202
MZF1_01	-292	-	0.968	8	CATGGGGA	complement of 210-217
ELK1_02	-105	+	0.963	14	CTCTCCGGAAGCCT	397-410
CETS1P54_01	-102	+	0.974	10	TCCGGAAGCC	400-409
AP1_Q4	-42	-	0.963	11	AGTGACTGAAC	complement of 460-470
AP1FJ_Q2	-42	-	0.961	11	AGTGACTGAAC	complement of 460-470
PADS_C	45	+	1.000	9	TGTGGTCTC	547-555

FIGURE 5